DOCUMENT RESUME

ED 265 224 TM 860 091

AUTHOR Herman, Joan
TITLE Report on the Revision of the CSE Evaluation Kit. Research into Practice Project.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Dec 85
GRANT NIE-G-83-0001
NOTE 255p.; For related documents, see ED 058 673 and ED 175 887-94.
PUB TYPE Reports - Descriptive (141) -- Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC11 Plus Postage.
DESCRIPTORS Achievement Tests; Attitude Measures; *Educational Assessment; *Evaluation Methods; Evaluators; *Formative Evaluation; *Program Evaluation; Program Implementation; Statistical Analysis; *Summative Evaluation
IDENTIFIERS CSE Evaluation Kit; National Institute of Education; *Research into Practice Project

ABSTRACT

This document describes the revision of the Center for the Study of Evaluation (CSE) Program Evaluation Kit, including planning, development, and/or revisions of specific components. The kit is a set of books, originally developed in 1978, and designed to provide step-by-step procedural guides to help people conduct evaluations of educational programs. To assure that the revision would accurately portray evaluation theory and state of the art practice, an advisory committee of five individuals was established. The committee met as an advisory board and agreed on the following changes to eight chapters: (1) Evaluation Handbook, provide orientation as to how the field has changed and emphasize on-going, internal improvement-oriented evaluation; (2) How to Deal with Goals and Objectives, change title and revise substantially; (3) How to Design a Program Evaluation, rewrite introduction; (4) How to Measure Program Implementation, add brief overview and expand discussion; (5) How to Measure Attitudes, essentially unchanged; (6) How to Measure Achievement; change title and expand discussion; (7) How to Calculate Statistics, essentially unchanged; and (8) How to Present an Evaluation Report, change title and expand. How to Conduct a Qualitative Study, a new addition to the kit, was recommended. Authors provided drafts, several of which are appended to this document. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Center for the Study of Evaluation

UCLA Graduate School of Education
Los Angeles, California 90024

ED265224

DELIVERABLE - DECEMBER 1985

RESEARCH INTO PRACTICE PROJECT

Report on the Revision of the CSE Evaluation Kit

TM 860 091

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C. Griffith
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

DELIVERABLE - DECEMBER 1985

RESEARCH INTO PRACTICE PROJECT

Report on the Revision of the CSE Evaluation Kit

Project Director: Joan Herman

Grant Number: NIE-G-83-0001

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was in part performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Report on the Revision of the CSE Evaluation Kit

CSE's Program Evaluation Kit is a set of books providing step-by-step procedural guides to help people to conduct evaluations of educational programs. Originally developed under a grant with the National Institute of Education and copyrighted in 1978, the Program Evaluation Kit is published by Sage Publications and includes the following eight books:

1. The Evaluator's Handbook serves as a organizing framework for the entire kit, taking the potential evaluator step-by-step through a generic procedure for conducting formative and summative evaluations. It also provides a directory to the rest of the kit. The introduction in chapter one calls attention to critical issues in program evaluation. Chapter 2, How to Play the Role of Formative Evaluator, describes the diversified job responsibilities of this role. Chapters 3, 4, and 5 contain step-by-step guides for organizing and accomplishing three types of evaluations:

   ° A formative evaluation calling for a close working relationship with the staff during program installation and development (Chapter 3)

   ° A standard summative evaluation based on measurement of achievement, attitudes, and/or program implementation (Chapter 4)

   ° A small experiment, a procedure most likely to be of interest to a researcher or to the evaluator who wishes to either conduct pilot tests or evaluate a program aimed toward a few measurable objectives (Chapter 5)

The Handbook concludes with a Master Index to topics discussed throughout the Kit.

2. How To Deal With Goals and Objectives provides advice about using goals and objectives as methods for gathering opinions about what a program

should accomplish. The book then describes how to organize the evaluation around them. It suggests ways to find or write goals and objectives, reconcile objectives with standardized tests, and assign priorities to objectives.

3. How To Design a Program Evaluation discusses the logic underlying the use of research designs - including the ubiquitous pretest-posttest design - and supplies step-by-step procedures for setting up experimental, quasiexperimental, and time series designs to underpin the collection of evaluation data. Six designs, including some unorthodox ones, are discussed in detail. The book outlines the use of each design, from initial choice of program participants to analysis and presentation of results. Finally, it includes instructions about how to construct random samples.

4. How To Measure Program Implementation presents step-by-step methods for designing and using measurement instruments - examination of program records, observations, and self-reports - to accurately describe how a program looks in operation. The first chapter discusses why measuring implementation is important and suggests several points of view from which you might describe implementation, for instance, scrutinizing the consistency of the program with what was planned or writing a naturalistic description free of such preconditions. Its second chapter is an outline of the implementation section of an evaluation report.

5. How To Measure Attitudes should help the evaluator select or design credible instruments for attitude measurement. The book discusses problems involved in measuring attitudes - including peoples' sensitivity about this kind of measurement and the difficulty of establishing the reliability and validity of individual measures. It lists myriad sources of available

attitude instruments and gives step-by-step instructions for developing questionnaires, interviews, attitude rating scales, sociometric instruments, and observations schedules. Finally, it suggests how to analyze and report results from attitude measures.

6. How To Measure Achievement focuses primarily on the tests administered for program evaluation. The book can be used in several ways. In case you plan to purchase a test, it helps you find a published test to fit your evaluation. To this effect, the book lists anthologies and evaluations of existing norm- and criterion-referenced tests and supplies a Table for Program-Test Comparison. The step-by-step procedure for completing this table directs you to compute numerical indices of the match between a particular test and the objectives of a program. If you want to construct your own achievement test, the book presents an annotated guide to the vast literature on test construction. Chapter 4 lists, as well, test item banks and test development and scoring services. The final chapter describes how to analyze and present achievement data to answer commonly-asked evaluation questions.

7. How To Calculate Statistics is divided into three sections, each dealing with an important function that statistics serves in evaluation: summarizing scores through measures of central tendency and variability, testing for the significance of differences found among performances of groups, and correlation. Detailed worksheets, ordinary language explanations, and practical examples accompany each step-by-step statistical procedure.

8. How To Present an Evaluation Report is designed to help you convey to various audiences the information that has been collected during the course of the evaluation. It contains an outline of a standard evaluation report;

directions and hints for formal and informal, written and oral, reporting; and model tables and graphs, collected from the Kit's design and measurement books, for displaying and explaining data.

Over the years, the kit has been widely used in the field and has been an important resource for training evaluators and for helping those charged with evaluation responsibilities to complete their tasks. In fact, over 150,000 units of the kit have been sold since it was first published. Although the kit continues to be distributed and to provide service, it has been over ten years since it was first developed, and during that time the field of evaluation has matured and changed considerably.

So too, the CSE evaluation kit needed to be changed to reflect these many changes and to continue to provide an updated, easy to follow resource for evaluation practitioners, a conclusion reached by jointly by CSE, its advisors, and Sage Publications. Consequently, CSE requested and received permission from the NIE to use resources from the Research Into Practice Project in partial support of the revision effort. This document describes the efforts supported with NIE funds, including planning for the revision and the development and/or revision of specific components.

Planning for the revision

A important component of the planning effort was the establishment of an advisory committee to assure that the revision would accurately portray evaluation theory and state of the art practice as it has evolved over the last fifteen years. Five individuals were approached and agreed to serve in this advisory role. Each of these individuals has played a prominent role both in conceptualizing the guiding theories and methodologies for the field and in the conduct of evaluation practice. They include:

° Robert Boruch, Northwestern University

° Ernie House, University of Illinois

° Gene Glass, University of Colorado

° Michael Patton, University of Minnesota

° Carol Weiss, University of Arizona

An advisory board meeting was convened to discuss potential revisions to the kit. Consensus was researched on the need for the following changes:

1. Evaluation Handbook: Provide orientation to how the field of evaluation has changed over the last 10 years, moving from an emphasis on mandated, external evaluations of federal and state-supported programs to concern with on-going, internal improvement-oriented evaluations of local programs; from concentration on experimental, quantitative methods to a consideration of more qualitative approaches; from near exclusive build-in sensitivity to stakeholders and potential utilization throughout the evaluation process, and to the influence of political and ethical issues.

2. How to Deal with Goals and Objectives: Change title to "How to Define Evaluation Goals" or "How to Define Your Role As Evaluator." A shortened, substantially revised version of the current test would constitute only one part of the new book; other concerns to be dealt with would include considering different potential purposes for the evaluation; framing evaluating questions based on client needs and theoretical predispositions; identifying and gathering input from a variety of stakeholders; setting priorities, etc.

3. How to Design a Program Evaluation: Rewrite introduction to the emergence and credibility of more qualitative approaches; refer reader to

qualitative methods book for design considerations for the qualitative context.

4. How to Measure Program Implementation: Add brief overview of the naturalistic research paradigm early in the book, spelling out the definitions of and differences between such things as naturalistic research, qualitative research, ethnographic research, responsive evaluation, etc.; refocus and enlarge current characterization of naturalistic/responsive observation, giving attention to important design issues (e.g. a priori vs. a posteriori design); expand discussion of observation and of focused interviewing; include discussion of data reduction and analysis.

5. How to Measure Attitudes: Essentially OK as is; examples outside of education may be helful; need external review.

6. How to Measure Achievement: Change title to "How to Measure Performance." Current discussion of achievement measures would be expanded to include performance measures and indicators in education and other fields. Test would consider issues such as deciding what to measure; sources and type of measures; selection criteria; procedures for constructing measures.

7. How to Calculate Statistics: Essentially OK as is; may want to consider adding more complex analysis strategies.

8. How to Present an Evaluation Report: Change to "How to Communicate Evaluation Findings" or "How to Report to Decisionmakers." In addition to dealing with how to write a report at the end of the project that is sensitive to user needs, the book would be expanded to consider reporting and communication needs throughout the evaluation process. Included would

be identification of target audiences and their information needs, analysis of evaluation and program context, formulation of strategies and timing for communicating evaluation results.

9. How to conduct a Qualitative Study: A new addition to the kit series, to include general orientation to qualitative strategies, guidelines on when to use qualitative methods and step-by-step procedures for design, observation, interviewing, and analyses and syntheses of data.

Development of Revised Manuscripts

Potential authors to accomplish the changes identified above were identified and contacted to determine their availability and interest. As a result of these interactions, the following individuals agreed to take responsibilities as follows:

1. Evaluation Handbook - Joan Herman & Marv Alkin (Editors for entire revision)
2. How to Define Your Role as An Evaluator - Brian Stecher, ETS
3. How to Design a Program Evaluation - Joan Herman
4. How to Measure Program Implementation - Jean King, Tulane University
5. How to Measure Performance - Joan Herman
6. How to Communicate Evaluation Findings - Phyllis Jacobson, Fillmore Unified School District
7. How to Conduct A Qualitative Study - Michael Patton, University of Minnesota

Of these revisions, drafts of #1-4 and 6 were to be accomplished within the grant period, with principal emphasis on the Evaluation Handbook. Authors of each of these revisions were asked to prepare a detailed outline. After these outlines were reviewed by the series editors and modified as necessary, authors started their writing tasks. Drafts of these manuscripts are appended in the following sections. Please note that the publisher (SAGE) is not requiring new camera-ready copy for the revisions. Therefore, in the interests of economy, zeroxed copies of portions of the current kit are included within the appended manuscripts.

EVALUATION HANDBOOK

(DRAFT)

Revision Editors: Joan Herman
Marv Alkin

December 6, 1985

TABLE OF CONTENTS

INTRODUCTION

- Components of the kit
- Kit vocabulary

CHAPTER ONE ESTABLISHING THE PARAMETERS OF AN EVALUATION

- An Evaluation Framework
- Conceptualizing the Evaluation
- How to Determine a General Technical Approach to an Evaluation
- How does an Evaluator Decide what to Measure?

CHAPTER TWO HOW TO PLAY THE ROLE OF AN EVALUATOR:

- A Review of Formative and Summative Evaluations

- Agenda A: Setting Boundaries for the Evaluation
- Agenda B: Selecting Appropriate Design and Measurements
- Agenda C: Data Collection and Analysis
- Agenda D: Final Reports

CHAPTER THREE STEP-BY-STEP GUIDES FOR CONDUCTING AN EVALUATION

- Agenda A
- Agenda B
- Agenda C
- Agenda D

CHAPTER FOUR STEP-BY-STEP GUIDE FOR CONDUCTING A SMALL EXPERIMENT

APPENDIX A: INDEX TO THE PROGRAM EVALUATION KIT

INTRODUCTION

The Program Evaluation Kit is a set of books intended to assist people who are conducting program evaluations. Its potential use is broad. The kit may be an aid both to experienced evaluators and to those who are encountering program evaluation for the first time. Each book contains step-by-step procedural guides to help people gather, analyze, and interpret information for almost any purpose, whether it be to survey attitudes, observe a program in action, or measure outcomes in an elaborate evaluation of a multi-faceted program. Examples are drawn from educational, social service, and business settings.

In addition to suggesting step-by-step procedures, the kit also explains concepts and vocabulary common to evaluation, making the kit useful for training or staff development.

COMPONENTS OF THE KIT

The Program Evaluation Kit consists of the following nine books, each of which may be used independently of the others.

1. This, The Evaluator's Handbook, provides an overview of evalulation activities and a directory to the rest of the kit. Chapter 1 suggests an evaluation framework which is based upon common phases of program development. Chapter 2 discusses things to consider when trying to establish the parameters of an evalation. Chapter 3 presents specific procedural agendas for conducting evaluations. Chapters 4, 5, and 6 contain specific guides for accomplishing three general types of evaluations: A formative evaluation, a standard summative evaluation, and a small experiment.

The Handbook concludes with a Master Index to topics discussed throughout the Kit.

2. How to Define Your Role as an Evaluator provides advice about focusing an evaluation, that is, deciding upon the major questions the evaluation is intended to answer and identifying the principal audience for the evaluation.

3. How to Design a Program Evalulation discusses the logic underlying the use of research designs -- including the uiquitous pretest-posttest design-- and supplies step-by-step procedures for setting up and interpreting the results from experimental, quasi-experimental, and time series designs. Six designs, including some unorthodox ones, are discussed in detail. Finally, the book includes instructions about how to construct random samples.

4. How to Use Qualitative Methods in Program Evaluation explains the basic assumptions underlying qualitative procedures and suggests the most appropriate situations for using qualitative designs in evaluations. (To be revised upon review of Michael Patton's manuscript).

5. How to Measure Program Implementation presents methods for designing and using measurement instruments --examination of program records, observations, and self-reports -- to accurately describe how a program looks in operation. (To be revised upon completion of Jean King's manuscript.

6. How to Measure Attitudes should help an evaluator select or design credible instruments to measure attitudes. The book discusses problems involved in measuring attitudes, including peoples' sensitivity about

this kind of measurement and the difficulty of establishing the reliability and validity of individual measures. It lists myriad sources of available attitude instruments and gives precise instructions for developing questionnaires, interviews, attitude rating scales, sociometric instruments, and observation schedules. Finally, it suggests how to anlayze and report results from attitude measures.

7. How to Measure Performance (complete the descsription upon review of the manuscript).

8. How to Calculate Statistics is divided into three sections, each dealing with an importnat function that statistics serves in evaluation: summarizing scores through measures of central tendency and variability, testing for the significance of differences found among performances of groups, and correlation. Detailed worksheets, non-technical explanations, and practical examples accompany each statistical procedure. (This may be revised based upon Gene Glass's suggestions.)

9. How to Present Evaluation Findings is designed to help evaluator convey to various audiences the information that has been collectaed during the course of the evaluation. It contains an outline of a standard evaluation report, directions for written and oral reporting, and model tables and graphs.

KIT VOCABULARY

For those who have had little experience with evalulation, it might be helpful to review a few basic terms which are used repeatedly throughout the Program Evaluation Kit. A PROGRAM anything you try because you think it will have an effect. A program might be something tangible such as a set of curriculum materials or a procedure, like the distribution of financial aid or an arrangement of roles and responsibilities, such as the reshuffling of administrative staff. A program might be a new kind of scheduling, like a four day work week; or it might be a series of activities designed to improve workers attitudes about their jobs. A program is anything definable and repeatable.

When you EVALUATE a program, you systematically collect information about how the program operates about the effects it may be having and/or to answer other questions of interest sometimes the information collected is used to make decisions about the program, for example, how to improve it, whether to expand it, or whether to discontinue it. Sometimes evaluation information has only indirect influence on decisions, sometimes it is ignored altogether. Regardless of how it is ultimately used, program evaluation requires the collection of valid, credible information about a program in a manner that makes it potentially useful.

Generally an evaluation has a SPONSOR. This is the individual or the organization who requests the evaluation and usually pays for it. If the members of a school board request an evaluation, they are the sponsors. If a federal agency requires an evaluation, the agency is the sponsor.

Evaluations always have AUDIENCES. An evaluation's findings are of course reported to sponsors, but there might be other people interested n or directly affected by the findings. A common audience for information collecterd during program develoment might consist of program planners, managers, and staff who run the program. Another audience might be the recipients of the services or products; for example, students, parents, or customers. If the program will be expanded to additional sites, or if it is reported in widely circulated publications, then the broader scientific, educational, public service or business community comprises an evaluation audience. In short, audiences are the groups that you will have to keep in mind as you conduct the evaluation. If your audiences share a common point-of-view about the program or are likely to find the same evaluation information credible, consider yourself lucky. This is not always the case.

For some evaluations, of course, the roles of evaluator, sponsor, and audience are all played by the same people. If teachers or managers decide to evaluate their own programs they will be at once the sponsors, the audience, the program managers, and the evaluators. Although the kit treats these roles as distinct, it is understood that people sometimes fill overlapping functions.

One decision that an evluator makes affects the credibility of the evaluation for many audiences. This is the selection of an EVALUATION DESIGN, a plan determining what individuals or groups will participate in the evaluation, what types of data will be collected and when evaluation instruments or measures will be administered and to whom. The instruments could include tests, questionnaires, observations, interviews, inspections of records, etc.) The design provides a basis for better understanding the program and its effects. More traditional quantitative designs focus primarily on measuring program results and comparing them to a standard. Such comparisons (including other programs) give some perspective about the magnitude of a program's effect and helps the evaluator and relevant audiences determine whether it indeed is the program which brings about particular outcomes. In contrast, newer qualitative designs focus on describing the program in depth and on better understanding the meaning and nature of its operations and effects.

The focus of the Program Evaluation Kit is the collection, analysis, and reporting of valid, credible information which can have some constructive impact on program decisionmaking.

CHAPTER I

ESTABLISHING THE PARAMETERS OF AN EVALUATION

AN EVALUATION FRAMEWORK

Literature in the changing field of program evaluation has been marked by various evaluation models which serve to conceptualize the field and to draw boundaries on the evalutor's role. Descriptions of some of the more prominent educational evaluation models appear in Table 1. If you plan to spend considerable time working as an evaluator, the references in this table and the readings listed in the For Further Reading section at the end of the chapter should help you catch up on what evaluators have said about their craft.

This kit has drawn its prescriptions about how to conduct program evaluations from most of the models in Table 1. Each model is appropriate to a particular set of circumstances and since the kit's purpose is to help you decide what to do in different situations, its advice is eclectic, borrowed from various models.

Most of the evaluation models described in Table 1 outline the technical procedures their proponents believe should be followed in evaluations; some also consider the socio-political factors which need to be considered. The Program Evaluation Kit also shows this dual concern. It focuses not only on how to accomplish the technical requirements of an evaluation but also on how to structure the evaluation to facilitate the use and impact of its findings. This pragmatic perspective reflects a common observation that evaluation since the early 1960's has been grossly underutilized.

Early evaluation models reflected a general optimism that systematic, scientific procedures would deliver unequivocal evidence of program success or failure. "Hard data" could provide both sound information for planning more effective programs and a rational basis for educational decisionmaking. It was assumed that clear cause-effect relationships could be established between programs and their outcomes and that program variables could be manipulated to reach desired effects. In light of these hopes, thousands of evaluations were conducted throughout the 1960's and 1970's. Unfortunately, most of these evaluations did not have the expected impact, and many have questioned whether these evaluations had any impact at all. Believing in the potential contribution of their work to educational planning and policy, evaluators became concerned about how to have their findings used and not simply filed. At the same time, they came to realize that all social programs are not discrete entities with easily recognizable stages in a predetermined process of natural development. Programs are often amorphous, complex mobilizations of human activities and resources, embedded in political and social networks. It is a rare program which exists in hermetically sealed isolation, perfectly appropriate for scientific measurement and duplication.

The Program Evaluation Kit reflects the need for a flexible approach that considers the complex environment in which a program exists as well as the purpose and context of its evaluation. An evaluator must be aware of the decision-making context within which the evaluation is to occur. She must consider the perceptions and expectations of various audiences, the developmental phase of the program under investigation, as well as the technical issue of which methodology to use in gathering data.

Evaluations are quite situation specific, but some generalizations or rules of thumb can be offered about how to conduct them. The following sections will explain four general phases during the life of a program when evaluations are commonly conducted. The phases are certainly not clearly separate. They quite often overlap, and some programs skip certain phases entirely. The evaluations during each phase differ according to their primary audiences, according to the decisions which sponsors will have to make, according to the timing of data collection and reporting, and according to the general relationship of the evaluator to the program during the course of the evaluation.

Program Initiation

Early in the development of a program, sponsors, managers, and planners consider the goals they hope to accomplish through program activities, and identify the needs and/or problems that a program is supposed to redress. Formally or informally, every program, in fact, goes through some kind of needs assessment even though it may not be obvious whose needs are being defined nor that the process is very rigorous. In some cases, the needs are simply assumed, and planners proceed to structure activities accordingly. At other times, the sponsor or funding agency more or less declares a need by making money available for programs aimed at general goals. Sometimes, however, a systematic effort is made to verify that perceived needs actually exist, to priortize their importance and/or to identify specific underlying problems. If a school program is intended as a response to community needs, for example, and evaluator conducting a needs assessment may gather information broadly from parents, teachers, students, and a sample of the broader community. Similarly in trying to help structure a program to increase staff morale, an evaluator may observe closely and broadly survey employees, their supervisors, experts, and others in order to uncover the source of the morale problem and its potential solution. Such formal needs assessments often try to gather input from a broad range of sources. Sometimes, however, a more restricted approach is preferable, e.g., where needs are very specialized or highly technical. In such cases, an evaluator might only solicit the opinions of experts.

The point is that programs are often initiated in response to critical needs, to achieve high priority goals, or to solve existing problems.

Program Planning

A second phase in the life of a program is its planning. Ideally, a program is designed to meet the highest priority goals established by a needs assessment. At times, the need to reach certain goals will prompt planners to design a new program from scratch, putting together materials, activities, and administrative arrangements that have not been tried before. Other situations will require that they purchase and install, or moderately adapt an already existing program. Both situations qualify as program planning -- something that has not occurred previously in the setting is created for the purose of meeting desired goals. During this phase, controlled pilot testing and field tests can be used to determine the effectiveness and feasibility of alternative methods of addressing primary needs and goals. While it is desirable, at this point, to establish plans for conducting evaluations, practice rarely meets this ideal.

Program Implementation

The third phase occurs as the program is being installed. Suppose, for example, that urban planners want to try out a new management information system. Purchases are made, boxes delivered, and training planned. This will be the first year of the new system. Ideally, the program's sponsors should give the new system a chance to make mistakes, solve problems, and reach the point where it is running smoothly before they decide how good or bad it is. All the time a program is in this implementation stage, subject to trial and error, the staff is trying to operationalize it suitably and revise it as necessary to meet their particular situation. Evaluations during this phase need to be formative, that is, seeking to describe how the program is operating and to suggest ways to improve it. Formative evaluation can take many forms such as special surveys of program services, enthnographic studies, or analyses of administrative records to determine how the program actually operates. In this formative case, the evaluator may work very closely with the program staff and report both formally and informally about findings as they emerge.

Program Accountability

When a program has become established with a permanent budget and an organizational niche, it might be time to question its overall impact. Judgments may need to be made about whether or not to continue the program, whether it should be expanded, and whether it might be used to other sites. During this phase, the evaluation is summative and it is of most direct concern to program policy makers.

Ideally, because the summative evaluator represents the interests of the sponsor and the broader community, he should try not to interfere with program operations. The summative evaluator's function is not to work with the staff and suggest improvements while the program is running, but rather to collect data and write a summary report showing what the program looks like and what has been achieved. Such ideal detachment, however, is rarely possible and even the most detached findings can serve a useful formative purpose. Evaluators, in fact, are often expected to serve both formative and summative functions, seeking to contribute to program improvement and to provide summary judgements of program worth. Such expectations must be approached cautiously as some objectivity may be lost if a summative evaluator scrutinizes a program in which she has developed a personal stake.

In recent years, organizations have turned more and more toward hiring evaluators who are permanent members of the staff. These internal evaluators generally perform formative functions. They are often an adjunct to management, working to increase organizational efficiency and effectiveness on a regular basis. However, care must be taken that the evaluator, regardless of where or how he is employed, maintain integrity, objectivity, and an appropriate sense of differentiation.

CONCEPTUALIZING THE EVALUATION

"We'd better have an evaluation of Program X," someone could decide and then appoint you to carry out that decision. Proceed with this caution:

Your first act in response to this assignment should be to find out what evaluation means in this instance. Find out what is expected. What information will the evaluation be expected to provide? Does the sponsor or another audience want more information than you can possibly provide? Do they want definitive statements that you will not be able to make? Do they want you to take on an agnostic or advocate role toward the program that you cannot in good conscience assume?

Failure to reach a common understanding about the exact nature of the evaluation could lead to wasted money and effort, frustration, and acrimony if sponsors feel they did not get what they expected. Step one in any evaluation is to negotiate!

Immediately after accepting the assignment, try to get a clear picture of what you will be expected to do. This conceptualization will have six major considerations, each negotiated with the sponsor and the audiences:

1. A decision about what people really want when they say they want an evaluation.
2. Identification of what the audiences will accept as credible information.
3. Choice of a reporting style. This may include the extent to which you report quantitative or qualitative information, whether you will write technical reports, brief notes, or confer with the staff, and the timing of important reports.
4. Determination of a general technical approach based upon information and credibility needs.
5. A decision of what to measure and/or gather information about.

6. Delineation of what you can accomplish within the constraints of the evaluation's budget and political situation.

Each of these six considerations will be discussed in more detail below.

Determining What People Really Want When They Say They Want an Evaluation

The sponsor who commissions the investigation might have in mind any one of several kinds of activities that could be called evaluations. They are all closely related, and in some cases more than one may be required for a single project. In general, the activities may be classified loosely into five types of evaluations based upon the ultimate use of the findings. A request for an evalution may actually be a charge to collect information:

- To conduct a needs assessment.
- To describe what a program looks like in operation. This is an implementation evaluation.
- To measure whether goals have been achieved.
- To help managers plan the program and keep it running smoothly. This is a formative evaluation.
- To help the sponsor and others in authority decide the program's ultimate fate. This is a summative evaluation.

Each of these activities requires a somewhat different approach and various amounts of time and money, so it is crucial that the sponsor's primary purposes for having the evaluation done be made as clear as possible. How will the findings be put to use? Who is the main audience? At what general stage in the development of the program is the evaluation taking place? Through frequent interactions with the sponsor early in the study, identify and focus the relevant evaluation questions.

The boxes on pages --- to --- describe the five kinds of investigations usually conducted under the title evaluation. Each is characterized by the types of questions which the sponsor and the evaluator might typicaly consider, by the general activities that could be expected to occur, and by the decisions that might be affected. Note that recommendations for conducting formative and summative evaluations encompass the activities required for needs assessment, program implementation evaluation, and assessment of goal achievement. The Program Evaluation Kit contains enough information to help you perform any of the five types.

CHART 1 NEEDS ASSESSMENT

(Also called an Organizational Review)

Significant questions for sponsors and evaluators:

*What are the goals of the organization?

*What should the program(s) try to accomplish?

Can goal priorities be determined?

Is there agreement on the goals from all groups?

To what extent are goals being met?

What in the organization is succeeding or failing?

Is there a need to establish new programs or to revise old ones based upon identified needs?

Activities:

The evaluator might discover that the aim of the evaluation is neither to decide between continuing or dropping a program nor to develop detailed activities to improve a program as it proceeds. Rather, a sponsor wants to discover problem areas in the current situation which might eventually be remedied. A needs assessment often precedes specific program planning and can be used to re-examine existing goals and/or to make implicit goals more explicit.

Decisions and actions likely to follow a needs assessment

The decisions following a needs assessment usually involve allocation of resources to meet high priority needs. New programs may be planned or old ones revised to address the identified needs. The survey of needs is itself the end product in this type of evaluation, unlike a formative evaluation, where the evaluator works with the organizational staff to improve identified weaknesses during the course of the investigation.

Kit components of greatest relevance:

How to Define Your Role as an Evaluator

How to Measure Achievement

How to Measure Attitudes

How to Measure Program Implementation

(To be revised upon revision of kit components)

CHART 2 PROGRAM IMPLEMENTATION EVALUATION

(Also called a Program Documentation)

Significant questions for sponsors and evaluators:

*What is happening in Program X?

*Is the program being implemented according to plan?

*What do participants in the program experience?

How many and which participants and staff are taking part?

What is a typical schedule of activities?

How are time, money, and personnel allocated?

How much does the program vary from one site to another?

Activities:

A description of program implementation focuses on the activities, materials, and administrative arrangements that comprise a program. It does not include an examination of the results of program activities as would a formative or a summative evaluation. The audience wants a description of who is doing what in the program or of how a requirement has been interpreted by the program planners and developers across sites. Be sure to make it clear that program activities will not be related to outcomes. For many audiences, a description of what is taking place is sufficient information for making decisions about the program. This is particularly true when the program is designed to reflect a philosophy of how organizations should be run in order to acheive long-term goals.

Decisions and actions likely to follow an implementation study:

The information from the evaluation may be included in a larger formative or summative investigation. Sponsors are likely to judge the program on the basis of whether or not they think the activities occurring are valuable in themselves or would probably be effective in achieving other goals.

Kit components of greatest relevance

How to Measure Program Implementation

How to Design a Qualitative Evaluation

(To be revised when the kit components are revised)

CHART 3 MEASUREMENT OF GOAL ACHIEVEMENT

Significant questions for sponsors and evaluators:

*Is Program X meeting its goals?

*Is the program meeting its goals?

How can goal attainment be measured most credibly?

Activities:

The evaluator attempts only to measure the extent to which the program's highest priority goals are being achieved. It is important to emphasize that you will not be able to state whether the program alone is responsible for the observed results and certainly not whether some other program would have been better. Even though looking at goal achievement alone usually provides a poor basis for judging a program's comparative merits, your results can still be of some use. Determining the extent to which achievement matches a set of carefully considered standards does give a basis for at least tentative conclusions about the program's quality.

Decisions and actions likely to follow measures of goal achievement:

Planners may choose to reconsider goals and to focus program activities more appropriately to achieve significant goals. The information from this type of evaluation might be used in a more extended formative or summative investigation.

Kit components of greatest relevance:

How to Measure Achievement

How to Measure Attitudes

(To be revised when kit components are revised)

CHART 4 FORMATIVE EVALUATION

Significant questions for sponsors and evaluators:

*How can the program be improved as it develops?

What are the program's goals and objectives?

What are the program's most important characteristics-materials, activities, administrative arrangements?

How are the program activities supposed to lead to attainment of the objectives?

Are the program's important characteristics being implemented?

Are they leading to achievement of the objectives?

What adjustments in the program might lead to the attainment of the objectives?

Which activities are best for each objective?

Are some better suited to certain participants?

What measures and designs could be recommended for use during a summative evaluation of the program?

Activities:

Formative evaluation encompasses the thousand-and one jobs connected with providing information for the staff to get the program running smoothly. It might even include conducting a needs assessment. Certainly it will involve some attention to monitoring program implementation and achievement of goals. In order to improve a program, it will be necessary to understand how well a program is moving toward its objectives so that changes can be made in the program's components. Formative evaluation is time-consuming because it requires becoming familiar with multiple aspects of a program and providing program personnel with information and insights to help them improve it. Before launching into formative evaluation, make sure that their is actually a chance of making changes for improvement - if no such possibility exists, formative evaluation is not appropriate.

Decisions and actions likely to follow a formative evaluation:

As a result of formative evaluation, revisions can be made in the materials, activities, and organization of the program. These adjustments are made throughout the course of the evaluation.

Kit components of greatest relevance:

All of them.

CHART 5 SUMMATIVE EVALUATION

Significant questions for sponsors and evaluators:

*Is Program X worth continuing or expanding, or should it be discontinued?

What are Program X's most important characteristics (materials, activities, administrative arrangements, etc.)?

Do the activities lead to goal achievement?

What programs are available as alternatives to Program X?

How effective is Program X in comparison with alternatives?

How costly is the program?

Activities:

The goal of summative evaluation is to collect and to present information needed for summary statements and judgments of the program and its value. The evaluator should try to provide a basis against which to compare the program's accomplishments. One might contrast the program's effects and costs with those produced by an alternative program that aims toward the same goals. In situations where such a comparison is not possible, participants' performance might be compared with a group receiving no such program at all. The standard for comparison might come from the norms of achievement tests or from a comparison of program results with the goals identified by the program designers of the community at large.

In some instances, summative evaluation is not appropriate. A summary statement should not be written, for instance, about a program that has not been in existence long enough to be fully developed. The more a program has clear and measurable goals and consistent replicable materials, organization, and activities, the more suited it is for a summative evaluation.

Decisions and actions likely to follow a summative evaluation:

Decision makers may use information from summative evaluations to help them decide whether to continue or to discontinue a program or whether to expand it or reduce it.

Kit components of greatest relevance:

All of them.

What Will Be Accepted as Credible?

In addition to finding out what your audiences want to know, you will need to discover what they will accept as credible information. The credibility of the evaluation will, of course, be influenced by your own credibility, a judgment that will be based on your perceived competence as well as your personal style. For some, your perceived competence in technical skills or in reporting may be most important. For others, your expertise in program subject matter may be a primary consideration. The audience will be less skeptical if they are confident you know what you are doing. A skilled evaluator also has excellent interpersonal skills and is able to nurture trust and rapport with various users and audiences.

Your audiences' willingness to accept without question what you report will be based on other criteria as well. For one thing, they will take account of your allegiances. An evaluator must be perceived as free to find fault -- whether or not she does. This means that you should not be constrained by friendship, professional relations, or the desire to receive future evaluation jobs. In addition, audiences will believe what the evaluator reports to the extent that they see her as representing themselves. The program staff, for instance, may be suspicious of a formative evaluator who will write a summary report at year's end to the funding agency. The agency, on the other hand, will read report suspiciously if it suspects that the evaluator's formative work has put her on "their side." Because of these credibility problems, evaluators with ambiguous formative-summative job descriptions have to arrive at a determination of their primary audience through negotiation.

Another determiner of how seriously the audience listens to your results is the method you use for gathering information. Methods of data gathering include the evaluation design; the instruments administered; the people selected for testing, questioning, or observation; and the times at which measurements are made. The specific methods you select will depend on whether you and your audience favor more quantitative or more qualitative approaches to the evaluation. When you choose a general approach, select instruments and designs, or construct a sampling plan, remember this: You cannot count on your audience to accept as credible the same sorts of evidence that you consider most acceptable. People are usually skeptical, for instance, of arguments they do not understand. You might have noticed that when reports filled with complicated data analysis are presented, people stay quiet until a few experts have given their opinions. Then everyone discusses the opinions of the experts.

Unless your audience expects complex analyses, you should keep your data gathering and analyses straightforward. Think of yourself as a surrogate, delegated to gather and digest information that your audience would gather on its own, were it able. Keep a few representative members of the evaluation; audience in mind, and ask yourself periodically: "Will Mr. Carson see the value in collecting this data or in doing this analysis?" A good way to find out, of course, is to ask Mr. Carson.

Remember, as well, that the general public tends to place more faith in opinions and anecdotes than do researchers -- at least usually. If you plan to collect a large amount of hard data, you will have to educate people about what it means.

Deciding on Reporting Requirements and Style

Why worry about reporting when you're conceptualizing your evaluation project? First, because each formal report requires time to write and produce, reporting can have important implications for the project budget, particularly if different reports are to be produced to meet the needs of different audiences. Formal reports, too, are only part of what is required to help assure a useful project: reporting, both formal and informal varieties, should be an on-going process during the life of an evaluation, not just an end of project product.

A second reason for considering reporting requirements early is their influence on the methodology and impact of the evaluation. How reports are perceived by their potential audiences, the credibility of the evidence presented and the persuasiveness of findings and conclusions is dependent not only on what is presented but on how it is presented as well. What kinds of reports are desired? Preferred reporting style is applicable here. It refers to the relative weight a report gives to quantitative and to qualitative data and the degree of formality with which the report is delivered. Do the report audiences, for example, prefer evidence in the form of tables of means, percentages, etc., in the form of charts and graphs, and/or in the form of characteristic anecdotes? Such preferences need to be articulated and negotiated early in the planning process. (The kit book, How to Communicate Evaluation Findings, should be of help to you as you negotiate a reporting strategy.)

Determining the General Evaluation Approach

Part of what determines the credibility of an evaluation, as mentioned above, is the credibility of the technical approach -- the credibility of the design, methods, measures etc. that are utilized for answering the questions of interest in the evaluation. How do you choose an appropriate technical approach? The answer lies in the interplay between the evaluator's predispositions, client preferences, and most importantly the information needs of the evaluation.

Quantitative Approaches

You are probably aware that technical approaches are often dichotomized into two general categories: Quantitative approaches and qualitative approaches. Quantitative approaches have been most prevalent historically in evaluation studies, particularly in evaluation studies intended to measure program effects. Quantitative approaches are concerned primarily with measuring a finite number pre-specified outcomes, with judging effects and with attributing cause by comparing the results of such measurements in various programs of interest, and with generalizing the results of the measurements and the comparisons to the population as a whole. The emphasis is on measuring, summarizing, aggregating and comparing measurements, and on deriving meaning from quantitative analyses. (Quantitative approaches also may be used in rating, classifying, and quantifying particular pre-defined aspects of program operations.) Such approaches utilize experimental designs, require the use of control groups, and are particularly important when program effectiveness is the primary evaluation issue.

The Importance of Design and Control Groups in Quantitative Approaches

Why are designs and control groups so important? You probably already know a bit about design -- that it involves assignment of students or classrooms to programs, and to comparison or control groups. The purpose of this discussion is to present you with the logic underlying the need for good design in evaluations where you want to show that there is a relationship between program activities and outcomes.

First consider the common before and after design. In the typical situation, a new program has been instituted and an evaluation planned. The evaluator administers a pretest at the beginning of the program, and at the end of the program, a posttest as in the following examples:

- A new district-wide mathematics program is evaluated. The California Achievement Test is administered in September and again in May.
- A new halfway house program has set itself the goal of decreasing recidivism in its juvenile clients. The evaluator observes and records the number of arrests and of convictions of its clients at the beginning of the year and then again at the end of the year.
- An objective of a corporate reorganization project is to increase staff morale and productivity. Staff fills in a questionnaire at the beginning of the year and then again at the end of the year; productivity indices likewise or computed at the beginning and end of the year.

Differences on the pre and posttests are then scrutinized to determine whether the program did what it was supposed to do. This is where the before-and-after design leaves the evaluation vulnerable to challenge. It fails to answer two important questions:

1. How good are the results? Could they have been better? Would they have been the same if the program had not been carried out?
2. Was it the program that brought about these results or was it something else?

Consider the folowing situtation. A new math program has been put into effect in the Lincoln District. Ms. Pryor, the superintendent, wants to assess the quality of the program by examining students' grade equivalent scores from a standardized math test given in September and then again in May. She notes that the sixth grade average was 5.4 in September and 6.5 in May. She attempts to judge the value of the new math program based on this pretest-posttest information.

Results on State Math Test - 6th Grades

Reading Program	Sept. Pretest (G.E.)	May Posttest (G.E.)
Sunnydate Learning Associates	5.4	6.5

The Lincoln students in the example have shown a considerable gain in reading from pretest to posttest - 1.1 grade equivalent points. On the other hand, they are still not performing at grade level. Therefore Ms. Pryor must ask herself, How good are these results? The answer depends, of course, on the children and the conditions in the school and home. For some groups, this would represent great progress; for others, it would indicate serious difficulties in the program.

How can Ms. Pryor find out what progress she should expect from her sixth graders? The pretest tells her something - the sixth-graders were six months behind in September, and they ended up only four months behind in May. Perhaps without the new program they would have ended up five months behind. Or perhaps they would have done better with the old program! In order to know what difference the program made, she needs to know how the students would have scored without the program.

Ms. Pryor has another problem in interpreting her results. She cannot even show that the gains she did get on the posttest were brought about by the new program. Perhaps there were other changes that occurred in the school or among the students this year - a drop in class size, or a larger number of parents volunteering to tutor, or the miraculous absence of "difficult" children who demanded teacher time and distracted the class. Many influences might cause the learning situation to alter from year to year.

Ms. Pryor could have ruled out most of these explanations of her results by using a control group. First, two randomly formed groups would have been assigned at the beginning of the year to either the new math program or to another semester of the old one (or to another alternative program). Before the program began, both groups would have been pretested. At the end of the year, the groups would be posttested using the same reading test.

Because the two groups were initially equivalent, the scores of the control group would show how the new program students would have scored if they had not received the new program:

Program	Pretest	Posttest
Sunnydate (x)	5.4	6.5
Old Program (Control Group)	5.4	6.1

But was it the new program that brought about the improvement, or was it some other factor? Using a true control group design, Ms. Pryor can discount the influence of other factors as long as these factors have probably also affected the control group. If, for instance, some students had had an enriched nursery school program that got them off to a good start in math, the random assignment should have spread these students fairly evenly between the two groups. If more parents were helping in the school, this should have benefitted both groups equally. If this year's sixth grade was generally quieter, with fewer difficult children, this should have affected both the experimental and control groups equally. Ms. Pryor does not even have to know what all the factors might have been. By randomly assigning the two groups, the influence of various factors affecting the math achievement of the two groups is likely to be equalized. Then differences observed in outcomes can be attributed to the one factor that has been made deliberately different: the reading program.

Though much maligned as impractical, the true control group design produces such credible and interpretable results that it should at least be considered an ideal to be approximated when evaluation studies are planned.[2] The design is valuable because it provides a comparative basis from which to examine the results of the program in question. It helps to rule out the challenges of potential skeptics that good attitudes or improved achievement were brought about by factors other than the program.

---

[2] Actually, true control group designs have been used in evaluation of many educational and social programs. A list of 141 of them, with references, is contained in Boruch, R.F. Bibliography: Illustrative randomized field experiments for program planning and evaluation. Evaluation, 1974, 2(1), 83-87.

It is not always easy to convince people that random assignment and experimentation are good things; and of course you must make decisions that are consistent with the opinions of your audience.

Consider using a design for planning the administration of each measurement instrument you will use. Consider a randomized design first. If this is not possible, then look for a non-equivalent control group - people as much like the program group as possible but who will receive no program or a different program. Or try to use a time-series design as a basis for comparison: find relevant data about the former performance of program groups or of past groups in the same setting. Only if none of these designs is possible should you abandom using a design. An evaluation that can say "Compared to such-and-such, this is what the program produced" is more interpretable than one like Ms. Pryor's that simply reports scores in a vacuum. (How to Design a Program Evaluation provides more detail on this subject.)

Qualitative Approaches Also Can Be Important

While experimental design and control groups have traditionally been advocated in evaluation studies, in recent years qualitative methods have been given increasing attention. In contrast to the traditional deductive approach used in quantitative approaches, qualitative methods are inductive. The researcher or evaluator strives to describe and understand the program or particular aspects of it as a whole. Rather than entering the study with a pre-existing set of expections or a prespecified classification system for examining or measuring program outcomes (and/or processes), the evaluator tries to understand the meaning of a program and its outcomes from the participants' perspectives. The emphasis is on detailed description and on in-depth understanding as it emerges from direct contact and experience with the program and its participants. Using more naturalistic methods of gathering data, qualitative methods rely on observations, interviews, case studies and other means of fieldwork. (*How to Use Qualitative Methods* provides more detail about qualitative approaches.)

Traditionally, qualitative and quantitative approaches have been seen as diametrically opposed, and many evaluators still strongly espouse one approach or the other. More recently, however, this view is beginning to change, and more and more evaluators are beginning to see the merits of combining both approaches in response to differing requirements within an evaluation and in response to different evaluation contexts. For example, if the purpose of an evaluation is to determine program effectiveness and the program and its outcomes are well defined, then a quantitative approach is appropriate. If, on the other hand, the purpose of an evaluation is to determine program effectiveness, but the program and its outcomes are ill-defined, the evaluator might start with a qualitative approach to identify critical program features and potential outcomes and then use a quantitative approach to assess their attainment. To take another, different example, suppose the purpose of an evaluation is program improvement, and more particularly to identify promising practices that might be updated in a number of program sites. An evaluator might use a quantitative approach to identify sites which were particularly successful in achieving program outcomes and then use a qualitative approach to understand how the successful sites were different from those with less success and to identify those practices which were related to their success.

There is no single correct approach, then, to all evaluation problems. Some require a quantitative approach; some require a qualitative approach; many can derive considerable benefit from a combination of the two.

How Does An Evaluator Decide What To Measure or Observe?

Having decided on a general approach, an evaluator might decide to measure, observe and/or analyze an infinite number of things: smiles per second, math achievement, time scheduled for reading, district innovativeness, sick days taken, self-concept, leadership, morale and on and on.

Carrying out an evaluation in any area often is a matter of collecting evidence to demonstrate the effects of a program or one of its subcomponents and/or to help improve it. The program's objectives, your role, and the audience's motives will help you to make gross decisions about what to look at. Four general aspects of a program might be examined as part of your evaluation:

- Context characteristics
- Student or Client characteristics
- Characteristics of program implementation
- Program outcomes
- Program costs

Context Characteristics

Programs take place within a setting or context - a framework of constraints within which a program must operate. They might include such things as class size, style of leadership in the school district organization, time frame within which the program must operate, or budget. It is especially important to get accurate information about aspects of the context that you suspect might affect the success of the program. If, for example, you suspect that programs like the one you are evaluating might be effective under one style of governance but not under another kind, you should try to assess leadership style at the various sites to explore that possibility.

Client Characteristics

Personal characteristics include such things as age, sex, socioeconomic status, language dominance, ability, attendance record, and attitudes. It may sometimes be important to see if a program shows different effects with different groups of clients. For example, if teachers say the least well-behaved students seem to like the program but the best behaved students do not like it, you would want to collect ratings of "well-behavedness" prior to the program and examine your results to detect whether these different reactions did indeed occur.

Characteristics of Program Implementation

Program characteristics are, of course, its principal materials, activities, and administrative arrangements involved in the program. Program characteristics are the things people do to try to achieve the program's goals. You will almost certainly need to describe these; though most programs have so many facets, you will have to narrow your focus to those that seem most important as most in need of attention. In summative evaluations, these will usually be the characteristics that distinguish the program from other similar ones.

Program Outcomes

You often will want to measure the extent to which goals have been achieved. You must make sure, however, that all the program's important objectives have been articulated. Be alert to detecting unspoken goals such as the one buried in this comment: "I could see how much the audience enjoyed the program. This alone convinced me the program was good." At least in the eyes of the person who said this, enjoyment was a program goal, or a highly valued outcome, whether or not this was so stated in program plans. You also need to ask whether outcomes are immediately measurable. Some hoped for outcomes may be so long-range that only a study of many years' duration could establish that they had occurred. This would be the case, for example, with goals such as "increased job satisfaction in adult life" or "a life-long love of books."

The evaluator should in general focus the evaluation on announced goals, but should be careful to include the possible wishes of the program's larger constituency - for example, the community - in formulating the yardsticks against which the program will be held accountable.

Program Costs

(insert to come)

Rules of Thumb

Beyond these general guidelines, decisions about exactly what information to collect will be situation-specific. Every program has distinctive goals; and every situation makes available unique kinds of data. Though there is no simple way to decide what specific information to collect, or what variables to look at, there are some rules of thumb you can follow:

1. Focus data collection where you are most likely to uncover program effects if any occur.
2. Try to collect a variety of information.
3. Try to think of clever - and credible - ways to detect achievement of program objectives.
4. Collect information to show that the program at least has done no harm.
5. Measure what you think members of the audience will look for when they receive your report.
6. Try to measure things that will advance the development of educational theory.

Use of each of these pointers is discussed below.

Focus Data Collection Where You Are Most Likely To Uncover Program Effects If Any Occur

While it is important that the evaluation in some way take note of major but perhaps ambitious or distant goals, do not place major emphasis upon them when deciding what to measure. One way to decide how to focus the evaluation is to classify program goals according to the time frame in which they can be expected to be achieved. Any particular intervention or program is more likely to demonstrate a detectable effect on close-in outcomes rather than those either logically or temporally remote. This means that you will also reduce the possibility of the program's showing effects if you focus on outcomes whose attainment is likely to be hampered by uncontrolled features of the situation. You should look for the program's effects close to the time of their teaching, and you should measure objectives that the program as implemented seeks to achieve.

Consider, for example, a hypothetical situation in which an employee training program has been designed with the objective of increasing the communication skills of employees working in programs for inner-city clients. The program was instituted in order to eventually accomplish these primary goals:

- To decrease employee absenteeism and early retirement because of high pressure on the job
- To encourage congenial interpersonal relationships among employees and clients
- To decrease the number of employee disciplinary referrals

In evaluating this program, you could measure the amount of employee absences and the number of hostile employee client encounters occurring before, during, and after the program; and the number of employees sent for disciplinary action. These, after all, are measures reflecting the program's impact on its major objectives.

There is a problem with basing the evaluation solely on these objectives, however. Judgments of the quality of the program will then be based only on the program's apparent effect on these outcomes. While these are the major outcomes of interest, they are remote effects likely to come about through a long chain of events which the employee training program has only begun. A better evaluation would include attention to whether employees learned anything from the training program itself or whether they displayed the behaviors the training was designed to produce.

In general, since there are various ways in which a program can affect its participants, one of the evaluator's most valuable contributions might be to determine at what level the program has had an effect. Think of a program as potentially affecting people in three different ways:

1. At minimum, it can make members of the target group aware that its services are available. Prospective participants can simply learn that the program is taking place and that an effort is being made to address their needs. In some situations, demonstrating that the target audience has been informed that the program is accessible to them might be important. This will be the case particularly with programs that rely on voluntary enrollment, such as life-long learning programs, a veneral disease education program, or community outreach programs for seniors, juveniles, etc. Evaluation of these kinds of programs will require a check on the quality of their publicity.
2. A program can impart useful information. It might be the case that a program's most valuable outcome is the conveyance of

information to some group. Learning, of course, is the major objective to most educational programs. Although most programs aim toward loftier goals than just the receipt of information, attention should not be diverted from assessing whether its less ambitious effects occurred. In the employee training example, for instance, it would be important to show that employees have become more aware of the problems and life experiences of minority clients. If you are unable to show an impact on their behavior, you can at least show that the program has taught them something.

3. A program can actually influence changes in behavior. The most difficult evaluation to undertake is one that looks for the influence of a program on people's day to day behavior. While behavior and attitude change are at the top of the list of many program objectives, actually determining whether such changes have occurred often requires more effort than the evaluator can muster. You will, of course, be interested in at least keeping tabs on whether the program is achieving some of its grander goals. Consider yourself warned, however, that the probability of a program showing a powerful behavioral effect might be minimal.

Try To Collect a Variety of Information

Three good strategies will help you do this. First, try to find useful information which is going to be collected anyhow. Find out which tests are given as part of the program or routinely in the setting; look at the teachers' plans for assessment; look at records from the program or at reports, journals, and logs which are to be kept. Check to see whether evidence of the achievement of some of the program's objectives can be inferred from these.

Another good way to increase the amount of information you collect is by finding someone to collect information for you. You might persuade teachers to establish record keeping systems that will benefit both your evaluation and their instruction. You might hire someone such as a student from a local high school or college to collect information. Perhaps you can even persuade a graduate student seeking a topic for a research study to choose one whose data collection will coincide with your evaluation.

Finally, a good way to increase the kinds of information you can collect is to use sampling procedures. They will cut down the time you must spend administering and interpreting any one measure. Choosing representative sites, events, or participants on which to focus, or randomly sampling groups for testing, will usually produce information as credible to your audiences as if you had looked at the entire population of people served.

Collecting a variety of information gives you the advantage of presenting a thorough look at the program. It also gives you a good chance of finding indicators of significant program effects and of collecting evidence to corroborate some of your shakier findings.

Besides accumulating a breadth of information about the program, you might decide to conduct case studies to lend your picture of the program greater depth and complexity. The case study evaluator, interested in the broad range of events and relationships which affect participants in the program, chooses to examine closely a particular case - that is, a school, a classroom, a particular group, or even an individual. This method enables you to present the proportionate influence of the program among the myriad other factors influencing the actions and feelings of the people under study. Case studies will give your audience a strong notion of the flavor of the activities which constituted the program and the way in which these activities fit into the daily experiences of participants.

Try To Think of Clever - and Credible - Ways To Detect Achievement of Program Objectives

Suppose in the teacher in-service example discussed earlier, it turns out that teacher absenteeism has remained unchanged and that the number of disciplinary referrals has diminished only slightly. These findings make the program look ineffective.

It might be the case, however, that though teachers have continued to send students to the office, they are discussing problems more often among themselves, reading more about minority groups, and talking more often with parents. Perhaps the content of referral slips has changed. Rather than noting a student's offense by a curt remark, maybe teachers are now sending diagnostic and suggestive information to the school office.

A little thought to the more mundane ways in which the program might affect participants could lead you to collect key information about program effects. A good way to uncover nonobvious but important indicators of program impact is to ask participants during the course of the evaluation about changes they have seen occurring. Where an informal report uncovers an intriguing effect, check the generality of this person's perception by means of a quick questionaire or a test to a sample of students. You should, incidentally, try to keep a little money in the evaluation budget to finance such ad hoc data gathering.

Collect Information To Show That The Program At Least Has Done No Harm

In deciding what to measure, keep in mind the possible objections of skeptics or of the program's critics. A common objection is that the time spent taking part in the program might have been better spent pursuing another activity. Sometimes the evaluation of a program, therefore, will need to attend to the issue of whether students or participants, by spending time in the program, may have missed some other important educational experience. This is likely to be the case with programs which remove students from the usual learning environment to take part in special activities. "Pull-out" programs of this kind are often directed toward students with special needs - either enrichment or remediation. You may need to show, for instance, that students who take part in a speech therpy program during reading time, have not suffered in their reading achievement. Similarly, you may need to show that an accelerated junior high school science program has not actually prevented students from learning science concepts usually taught at this level.

Related to the problem of demonstrating that students have not missed opportunities for learning is the requirement that you also show the program did no actual harm. For instance, attitude programs aimed at human relations skills or people's self-perceptions could conceivably go awry and provoke neuroses. Where your audience is likey to express concern about these matters, you should anticipate the concern by looking for these effects yourself.

Measure What You Think Members of the Audience Will Look For When They Receive Your Report

Try to get to know the audience who will recieve your evaluation information. Find out what they most want to know. Are they, for instance, more concerned about the proper implementation of the program than about its outcomes? A parent advisory group, for instance, might wish to see an open classroom functioning in the school. They may be more concerned with the installation of the program than with student achievement, at least during the first year of operation. In this case, your evaluation should pay more attention to measures of program implementation than to outcomes although progress reports will be appropriate as well. If you get to know your audience, you will realize that, for instance, Mr. Johnson on the school board always wants to know about integration or interpersonal understanding; or the foundation that supplied funding is mainly concerned with potential job skills. Visualize members of the audience reading or hearing your report; try to put yourself in their place. Think of the questions you would ask the evaluator if you were they.

Delineating What You Can Accomplish Within Budget and Other Constraints

Financial limitations and political climate represent important constraints on an evaluation, potentially limiting the scope and depth of its investigations. The amount of time an evaluator can devote, limitations on who, where, and when he can measure or observe, and constraints on what he can ask all determine the ultimate breadth and quality of an evaluation.

The amount of time an evaluator can devote to the project is dependent on the available budget. Available time, in turn, significantly influences methodological choices. Site visits, for example, are costly in terms of staff time as well as travel. Special outcome measures, as another example, requires staff time for development, pilot-testing and analysis. Assessing more rather than fewer program participants as a third example, has significant cost implications. Rarely are abundant resources available for an evaluation, and the evaluator often must juggle artfully to maintain a reasonable balance between the demands of scientific rigor and credibility and those of the budget. (Sometimes such a balance is just not possible and clients need to be informed accordingly.)

But financial resources represent only a part of the constraints on any evaluation. Some writers have expressed pessimism about the usefulness of evaluation results because of the overriding social and political motives of the people who are supposed to use evaluation results for making decisions, Ross and Cronbach[1] describe the situation this way:

> Far from supplying facts and figures to an economic man, the evaluator is furnishing arms to a combatant in a war with fluid lines of battle and transient alliances; whoever can use the evaluators to gain an inch of terrain can be expected to do so...The commissioning of an evaluation...is rarely the product of the inquiring scientific spirit; more often it is the expression of political forces.

---

[1] Ross, L. & Cronbach, L.J. "Review of the Handbook of Evaluation Research." Educational Researcher, 1976 5(10), 9-19.

The political situation could hamper an evaluation in several ways. For one, it might place constraints on data collection that make accurate description of the program impossible. The sponsor could, for instance, restrict the choice of sites for data collection, regulate the use of certain designs or tests, or withhold key information. Politics could, as well, cause the evaluator's report to be ignored or his results to be misinterpreted in support of someone's point of view.

Responding to any of these situations will depend on vigilance in each unique case. Remember that your major responsibility as an evaluator is to collect good information wherever possible.

How might an evaluator alleviate some of these political forces? First remember the old adage "Forewarned is forarmed" and be aware of the political forces at work in your situation. Second, try to neutralize the influence of competing agendas by drawing the representatives of powerful constituencies into the evaluation process. Identify the relevant decision makers and information users and work with them to identify the program needs and to focus the evaluation. On this point; The Standards for Evaluations of Educational Programs, Projects, and Materials developed by the Joint Committee on Standards for Educational Evaluation (1981) states:

> The evaluation should be planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation may be obtained, and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted.

Because of the acknowledged political nature of the evaluation process and the political climate in which it is conducted and used, it is imperative that you as the evaluator examine the circumstances of every evaluation situation and decide whether conforming to the press of the political context will violate your own ethics. It could turn out that the data that audiences want, or the kinds of reports required, do not suit your own talents or standards, or the standards of the profession.

The point is that all evaluations operate within a set of constraints -- financial, political, and others that influence both what an evaluation can accomplish and its potential impact. The evaluator needs to be aware of these various constraints and to plan accordingly for the most effective evaluation possible.

CHAPTER 2

HOW TO PLAY THE ROLE OF EVALUATOR

It is a rare evaluation that does not have both formative and summative characteristics. As the demand for utility in evaluation has increased over the years, information derived from summative evaluations is often used for some aspect of program improvement or renewal. In the ideal, a summative evaluation is designed primarily to assess the overall impact of a developed program so that decison-makers might determine the ultimate fate of a program, and a formative evaluation is conducted while a program is still being installed so that it may be implemented as effectively as possible. In practice, both types of evaluations pass through the same basic steps, although there may also be differences in timing, in audiences, and in the relationship between the evaluator and the program under investigation.

This chapter presents a description of the many facets of an evaluator's role and outlines the responsibilities and activities associated with the position. Since the tasks of an evaluator change with the context, it is inappropriate to prescribe what this person must do. Rather the chapter will describe the role with regard to the program and suggest some of the activities in which an evaluator might become involved. In this Handbook, the sets of tasks which evaluators aim to accomplish are called agendas. These are:

* Agenda A: Set the boundary of the evaluation.
* Agenda B: Select appropriate evaluation designs and

measurements.

- * Agenda C: Collect information and analyze data.
- * Agenda D: Report and confer with the primary audience(s).

You might think of each agenda as a set of information-gathering activities culminating in one or more meetings where decisions are made about the next information-gathering cycle. Although it would be logical to perform the tasks subsumed under each agenda in the sequence presented above, you will likely find yourself working at two or more agendas simultaneously or cycling back through them again and again. This will be particularly the case with Agendas C and D; you might collect, report, and discuss implementation and progress data many times during the course of the evaluation. Meetings with staff are also likely to involve more than one agenda.

This chapter discusses in detail various ways to accomplish each of these tasks. Chapter Three then presents step-by-step guides for completing each of the activities outlined here.

Special Focus of the Formative Evaluator

Whatever their situation, formative evaluators do share a set of common goals. Their major aim, of course, is to ensure that the program be implemented as effectively as possible. The formative evaluator watches over the program, alert both for problems and for good ideas that can be shared. The goal of bringing about modifications for a program's improvement carries with it four subgoals:

- To determine, in company with program planners and staff, what sorts of information about the program will be collected and shared and what decisions will be based on this information
- To assure that the program's goals and objectives, and the major characteristics of its implementation, have been well thought out and carefully recorded
- To collect data at program sites about what the program looks like in operation and about the program's effects on attitudes and achievement
- To report this information clearly and to help the staff plan related program modifications

2-2

Agenda A: Set the Boundaries of the Evaluation

Your first job will be to delineate the scope of the evaluation by sketching out with the program decision-makers or staff in a description of *what your tasks will be.* This first plan might result in a contract describing what you will do ~~for the staff~~, as well as the responsibility *they* will assume to help you gather information, and to act upon what you report.

As soon as you have hung up your telephone, having spoken with someone requesting ~~a formative~~ evaluation, you are confronted with Agenda A. It could be that the person asked you outright to help with project improvement. Or perhaps you *chose* to focus on providing formative information to the project after having been given the job of summative evaluator or an ambiguous evaluation role. It could be as well that you started out as one of the program's planners and that your role as formative evaluator is simply a "change of hats," a shift from your previous responsibilities.

Research the Program. You should find out as much as possible about the program before meeting with your audience, program staff and planners in the case of a formative evaluation and primary decision-makers in the case of a purely summative evaluation. A recommended first activity is to contact someone who is familiar with this or similar programs. In addition to sharing basic information, he may be able to help you anticipate problems with the evaluation or with developing a good relationship with ~~staff.~~ and management.

By all means ask the program planners for documents related to funding and development or adoption of the program. These documents might include an RFP (Request for Proposals) issued by the funding source when it first offered money for such programs, the program plan or proposal, and program descriptions written for other reasons such as public relations. Use these documents to form an initial general understanding of what the program is supposed to look like, what its goals might be, and particularly, what shape the evaluation might take.

In addition, it may be worth your while to quickly check the educational literature to see what, if anything, has recently been written about programs like the one in question, or about ~~specific components--say, commercial curriculum materials, that planners intend to use~~ the program [illegible] models. You may even find earlier evaluations of this or similar programs.

2-3

Encourage Cooperation. For whatever reason you undertake the evaluation, the first step will be to establish a working relationship with the clients. Since formative evaluation depends on sharing information informally, one of the outcomes of Agenda A in this type of evaluation should be the establishment of groundwork for a trusting relationship with the staff and planners. If your evaluation has been commissioned by the program staff itself, establishing trust will be easier than if, as in many summative evaluations, the contractor is an external agency, perhaps the state or federal government. In the latter case, the evaluator starts out as an outsider.

It is very important to avoid ending up in an adversary position against a defensive program staff. This is particularly important in a formative evaluation, and your posture toward the staff and the program will differ somewhat from that taken in a summative evaluation. A formative evaluator will need to convince the staff that her primary allegiance is to help them discover how to optimize program implementation and outcomes. Whereas, it is clear in a summative evaluation that your main responsibility is to provide an unbiased report of program accomplishments to primary decision-makers, and this responsibility should be made quite explicit.

In order to develop the necessary trust in a formative evaluation, you might describe the form that your outside reporting will take and allow the staff a chance to review your external reports. Whenever it seems necessary, you might also guarantee that information shared for the purpose of internal program review and improvement will be kept confidential. An important way to gain the confidence of program personnel is to make yourself useful from the very beginning, efficiently collecting information they need or would like to have.

Elicit Information from Staff. While mutual trust must be worked out gradually, some more practical aspects of your role can be negotiated during a single meeting. Agenda A requires that you and the principal contractor for the evaluation decide together what you will do for them.

Again, an evaluation done for formative purposes requires a close working relationship with the staff as you jointly determine the most appropriate course for your efforts. If you arrive early during program development, you may find that the staff needs help in identifying program goals and choosing related materials, activities, etc. Even after the program has begun, they may still be planning. Regardless of the state of program development when you begin, you should help the staff outline what they consider to be the primary characteristics of the program, highlighting those which they consider fixed and those which they consider *changeable enough to be the focus of formative evaluation.*

It is important to get a clear picture of the attitudes of participants ~~teachers~~ and planners, particularly concerning their *commitment to change,* that is, the extent to which they are willing to use the information you collect to make modifications in the program. Though neither you nor they will be able to anticipate beforehand precisely what actions will follow upon the information you report, you should get some idea of the extent to which the staff is willing to alter the program.

In general, laying the groundwork for a ~~your~~ formative evaluation means asking the planners and staff such questions as:

- Which parts of the program do you consider its most distinctive characteristics, those that make it unique among programs of its kind?
- Which aspects of the program do you think wield greatest influence in producing the attitudes or achievement the program is supposed to bring about?

- What components would you *like* the program to have which it does not contain currently? Might we try some of these on a temporary basis?
- Which parts of the program as it looks currently are most troublesome, most controversial, or most in need of vigilant attention?
- On *what* are you most and least willing, or constrained, to spend additional money? Would you be willing or *could* you, for instance, purchase another mathematics series? Can you hire additional personnel or consultants?
- Where would you be most agreeable to *cutbacks?* Can you, for instance, remove personnel? If ~~the audio-visual learning~~ expensive equipment were found to be ineffective, would you eliminate it? Which ~~books,~~ materials, and other program components would you be willing to delete? Would you be willing to scrap the program as it currently looks and start over?
- How much administrative or staff reorganization will the situation tolerate? Can you change people's roles? Can you add to staff, say, by bringing in volunteers? Can you move people ~~—teachers, even students—~~ from location to location permanently or temporarily? Can you reassign students to different programs or groups?
- How much ~~instructional and curricular~~ change will you tolerate in the program beyond its current state? Would you be willing to delete, add, or alter the program's objectives? To what extent would you be willing to change books, materials, and other program components? Are you willing to rewrite lessons?

The objective behind asking these questions is *not* to record a detailed description of the program. This will be done under Agenda B. Rather, the purpose is to uncover particularly maleable aspects of the program. The best way

BEST COPY

to find out about the staff's commitment to change is to ask these hard questions early. A dedicated staff that has worked diligently to plan the program will likely have in mind a point beyond which it will not go in making modifications. You should locate that point, and choose the program features you will monitor accordingly.

Another important consideration in uncovering staff loyalties and attitudes is their commitment to a particular *philosophy* ~~of education~~. If they are adopting a canned program, this philosophy probably motivated their choice. Staff members developing a program from scratch may also subscribe to a single motivating philosophy. However, you may find it poorly articulated or even unclearly evidenced in the program. In this case, you can create a basis for future decision-making by helping the staff to clarify and put into practice what their philosophy says.

If you can help the staff outline areas of the program where modifications are likely to be either necessary or possible, then they can begin to delineate the parts of the program whose effectiveness should be scrutinized. This will, in turn, suggest the kinds of information they will need. If the program is based on canned curricula which will simply be *installed*, or on materials not expressly designed for the type of program in question, then what you can change will be restricted. In this case you should focus on how best to make materials or procedures fit the context.

Example. A group of language arts teachers in a large high school decided that an alarming number of ninth-grade students were unable to read at a level sufficient to appreciate the literary content of their courses. They decided to institute a tutoring program in which twelfth graders would spend three forty-five minute periods a week reading literary selections with ninth graders. The aim of the project was to improve the ninth-graders' reading as well as to introduce them to English literature. The reading selections used for this program were from *Pathways to English*, a popular ninth-grade anthology; the district budget did not include funds to purchase reading materials for secondary students.

The school's assistant principal, Al Washington, monitored closely the progress of the tutoring program. After only three months had passed, he noticed some disappointment among the initially enthusiastic teachers. Their informal assessments of the reading of tutees had convinced them that little progress had been made. Although the students enjoyed the tutoring experience, they were not learning to read. The teachers asked Mr. Washington to evaluate the program with an eye toward suggesting changes in the materials or the tutoring arrangements that would help the ninth graders with their reading. Mr. Washington carefully examined the *Pathways to English* text, observed tutoring sessions, and interviewed tutors and tutees. From this information he drew three conclusions about the program and offered suggestions for remedial action: (1) the vocabulary in *Pathways to English* was too difficult for the ninth graders, so each unit should be preceded by a vocabulary drill using a standard procedure that would be taught to tutors; (2) ninth graders were *listening* more than reading, so tutoring sessions should be restructured to follow a "you read to me and I read to you" format in which twelfth and ninth graders alternate reading passages; (3) the program as constituted gave ninth graders no feedback about their progress in either reading or literary appreciation - therefore, the teachers should write short unit tests in vocabulary, comprehension, and appreciation.

In the case where a wholly new program is being developed, you will want to identify the most promising sorts of modifications that can be made within existing budget limitations. You may find it most useful to concentrate on helping the staff select from among several alternatives the most popular or effective form the program can take.

Example. KDKC, an educational television station serving a large city, received a contract from the federal government to produce 13 segments of a series about intercultural understanding directed at middle grade students. The objective of the series would be to promote appreciation of diverse cultures by depicting life in the home countries of the major cultural groups comprising the population of the United States.

The producers of the series set out at once to assemble the programs based on the format of popular primary grade programs: the central characters living in a culturally diverse neighborhood converse with each other about their respective backgrounds. These conversations lead into vignettes—filmed and animated—depicting life and culture in different countries. Some members of the production staff, however, suggested that a program format suitable for the primary grades may "bomb" with older students. "How do we know," they asked, "what interests 10- and 11-year olds?" They suggested two formats which might be more effective: a fast action adventure spy story with documentary interludes and a dramatic program focusing on teen-age students traveling in different countries.

To test these intriguing notions, the producer called on Dr. Schwartz, a professor of Child Development. Dr. Schwartz, however, had to admit that he was not sure what would most interest middle grade students either. Since the federal grant included funds for planning, Dr. Schwartz suggested that the producer assemble three pilot shows presenting basically the same knowledge via each of the three major formats being considered and then show these to students in the target age group, assessing what they learned and their enjoyment. The producer liked the idea of letting an experiment determine the form of the programs and agreed to allow Dr. Schwartz to conduct the studies, serving as a formative evaluator.

¶ Outline the services you can provide.

In accomplishing Agenda A, you will want to convey to the key decision makers and information users ~~project staff and planners~~ a description of what you can and cannot do for them within the constraints of your abilities, time, and budget. You should let them know the sorts of choices you will have to make based on staff preferences and likely future circumstances. It is also desirable that you frankly discuss both your areas of greatest competence and those in which you lack expertise. The staff should know in what ways you believe you can be of most benefit to the program as well as how the program might profit from the services of a consultant who could handle matters outside your competence.

Although you should have an evaluation plan in hand before you meet with staff and planners, let your audience have the opportunity to select from among several options; present your preferences as recommendations, and negotiate the general form your evaluation services will take. Try not to become enmeshed in details too early. You need only agree initially on an outline of your evaluation responsibilities. As the program develops, these plans could easily

2-7

change. When describing the service you might perform, list the kinds of questions you will try to answer about ~~academic progress~~ productivity, effective use of materials, proper and timely implementation of activities, adequacy of administrative procedures, and changes in attitudes. Describe, as well, the supporting data you will gather to back up depictions of program events and outcomes.

If you feel that the situation will accommodate the use of a particular evaluation design, then propose it and describe how designs increase the interpretability of data. In ~~cases~~ a formative evaluation where you note controversy over the inclusion of a program component, or where there exists a set of ~~instructional~~ alternatives without a persuasive reason to favor any one of them, suggest pilot studies based on *planned variation.* These studies, which could last just a few weeks, would introduce competing variations in the program at different sites. To help the planners eventually choose among them, you would check their ease of installation, their relative effect on ~~student~~ achievement, and staff and ~~student~~ client satisfaction.

Example. The curriculum office of a middle-sized school district had purchased an individualized math concepts program for the primary grades. The program materials for teaching sets, counting, numeration, and place value consisted of worksheets and workbooks and sets of blocks and cards. Curriculum developers familiar with the literature in early childhood education were concerned about the adequacy of the "manipulatives" for conveying important basic math concepts. They wondered, as well, whether the materials would maintain the interest of young children. To find out whether *supplementary* materials should be used, the Director of Curriculum set up a pilot test. She purchased some Montessori counting beads and cuisenaire rods from a commercial distributor and contacted a group of interested teachers to write supplementary lessons for using the beads and rods.

When the program began in September, most of the district's schools used the new program without supplementary materials. Randomly selected schools were assigned to receive the teacher-made lessons based on commercial manipulatives. An in-service workshop was held at the end of the summer to familiarize teachers in the pilot schools with the commercial materials and locally made lessons.

The Director of Curriculum periodically monitored the entire new program, administering a math-concepts test to representative classrooms three times during the first semester. When these tests were administered, she took special care to include in the sample the classrooms using the teacher made lessons. She was therefore able to use the classes without supplementary materials as a control group against which to measure student achievement. Since development of mathematical concepts is difficult to measure in young children, she also planned to monitor teacher estimates of the suitability, ease of installment, and apparent effectiveness of the various program versions.

Planned variation studies for a program under development from scratch might emphasize the relative effectiveness of different materials and activities. Where a previously designed program is being adapted to a new locale, planned variation studies will more likely look at variations in staffing and program management.

2-8

If there is enough time, suggest a balanced set of data collection activities. Especially in the case of a formative evaluation, include a few important pilot studies, continuous monitoring of program implementation, and periodic checks on achievement and attitudes. The precise details of these plans can be worked out under Agenda D. If possible, a formative evaluation should include at least one service to the program that requires your frequent presence at program sites and staff meetings. This will help you stay abreast of what is happening and maintain rapport with the staff. Such extensive contact with staff is generally not necessary for a summative evaluation.

Arrive at a contract

Once you and your clients have reached an agreement about your role and activities, write it down. This tentative scope of work statement should include:

- A description of the evaluation questions you will address
- The data collection you have planned, including sources, sites, and instruments
- A timeline for these activities, such as the one in Table 2, page 28
- A schedule of reports and meetings, including tentative agendas where possible

Be certain to stress the *tentative* nature of this outline, allowing for changes in the program and in the needs of the staff. Also, remember *you* will be responsible for all evaluation activities contracted. Exercise your option to accept or reject assignments.

The linking agent role in formative evaluation

If you have expertise in or access to information about the subject areas the program addresses or if you know about programs of its type in operation elsewhere, you might like to append to your formative role an additional title, much in vogue--*Linking Agent.* A linking agent *connects* important accumulated information and resources with interested parties, in this case, the planners and staff of the program you are evaluating. The linking agent is a one-person information retrieval system. Her sources are libraries, journals, books, technical reports, and experts and service agencies of all kinds.

Different linking services will be relevant to different programs. For example, you might locate and describe for the staff sets of recently developed curriculum materials related to a locally developed program. If you were evaluating a special education program, you might find and make use of a regional resource center offering consultation and diagnostic help with special education students. The role of linking agent will simply broaden the range of program improvement information you collect. Be careful, however, that linking does not interfere with your primary job—to monitor and describe the program at hand.

Agenda B: Select Appropriate Evaluation Designs and Measurements

In Agenda A, you committed yourself to evaluation activities. In this agenda, the program's decision-makers, staff provide a working description of the program. In a summative evaluation, you will collect a statement of the program's goals and objectives, a description of how the program components have been implemented, and a summary of the costs of the program in order to decide which outcomes, activities, and costs to measure. The kit book *How to Design a Program Evaluation* will give you careful guidance about selecting the appropriate design for the evaluation. A design is a plan of which groups will take part in the evaluation and when measurements will be made on these groups. Your design will might include a wide variety of measures such as achievement assessments, attitude scales, narrative descriptions of observations, and cost analyses. All ~~of~~ measures should be carefully selected to give information about particular outcomes.

Prepare a Program Statement. If a program is still under development, then it may be your task to have the program's planners and staff commit themselves to a working description of their program. The final product of this activity should be a written list of ⟶

(to next page)

TABLE 2
A Summary of the
Formative Evaluator's Responsibilities

Tasks/Activities	Time in Months 19XX J J A S O N D	19YY J F M A M	Completion Date	Reports and Deliverables	Number of personnel work hours consumed: Program Evaluator	Program Director	Teaching Staff	Principals	Teacher Aides	Clerical Staff
Review/revision of program plan	—		July 31	Revised written plan	37	8	-	6	-	16
Discussion about method of formative feedback alternatives	—		Sept 15	None	16	7	24	6	-	-
Planning of implementation-monitoring activities	—		Sept 30	List of instruments; Schedules of classroom visits	60	10	24	-	-	2
Construction of implementation instruments	—		Oct 10	Completed instruments	60	5	12	-	-	16
Planning of unit tests	—		Oct 10	List and schedule of achievement tests	30	5	-	-	-	2
First meeting with staff	¦		Nov 1	None	9	15	24	2	20	-
First meeting with district administration	¦		Nov 8	First interim report	22	20	-	-	-	30
				TOTAL PERSON HOURS						

program objectives, and a *rationale* that describes the relationship between these objectives and the activities that are supposed to produce them. The program statement should reflect the current consensus about what comprises the program arrived at with the understanding that the program's character may alter over time. ~~Although the scope of activities in Agenda B is relatively restricted, you will find that the ease of the task will depend heavily on your interpersonal skills.~~

Writing a preliminary program statement—even if only in outline form—is useful because it demands careful thought by the program's staff and planners about what they intend the program to look like and do. This thinking alone can lead to program improvements. Most successful programs are built upon a structured plan that has been clearly thought out and that describes as precisely as possible the program's activities, materials, and administrative arrangements. A clear program statement encourages program success for several reasons. For one thing, everybody involved knows where such a program is headed and what its critical characteristics ought to be. Everybody is working from the same plan. If program variations are taking place, then the staff and planners are likely to be aware of this. The evaluator can document such differences and, where possible, assess their merits. Fear of disputes among staff members, advisory committees, and teachers should not dissuade you from attempting to clarify program procedures and goals. Disagreements during the planning of a program are by and large healthy, especially when a program is in its *formative* stage, and the staff should be willing to adopt a "wait and see" attitude. Not all differences of opinion may be resolved, but the pooling of staff intelligence through discussion should be preferred to leaving each teacher to make his own guesses about what will work best.

Make sure that goals are well stated

There are three basic sources of information ~~for Agenda B~~ about program goals:

- The program plan, proposal, and other official documents
- Structured interviews and informal dialogues with program staff
- Naturalistic observation-based intuitions about program emphases

As has been mentioned, it is possible that you will arrive on the scene and find that the program has been too vaguely planned. Formative evaluation presumes the legitimacy of evaluating programs whose content and processes are still developing. Frequently the staff will be unable to tell you exactly what the program should look like, and objectives may be too general to serve as a basis for monitoring pupil progress. Although you should get a glimpse of how the program will function from documents such as the program's proposal, or the program plan, often these consist of exhaustive lists of documented *needs* that the program should meet, a page or two on objectives, and a description of the program's staffing and budget. A description of what people taking part in the program do or have done to them is not to be found.

Official documents represent *formal* statements of program intentions. These may be outdated, incomplete, erroneous, or unrealistic. Written descriptions of categorically funded programs ~~such as ESEA Title I~~ are particularly misleading: their objectives often reflect only politically minded rhetoric. Canned programs, or sets of published program materials, are another source of official objectives. But be careful here as well. While adoption of a particular program may reflect a philosophy shared between program staff and the developer of the materials, it is also possible that the staff running this particular program consciously or unconsciously posseses a different set of goals or that the program will only use certain components of the purchased materials.

Because of the problems associated with goals listed in official documents, you are responsible for obtaining goal information from discussions which probe the motives of the program staff and from observations of the program. Simply *asking* staff members their perceptions of program objectives will often elicit a recitation of documented goals, cliches, or socially desirable answers. Asking staff for *scenarios* of what you might see or expect to see at program sites is sometimes more productive. These scenarios can be followed by questions about the particular learning that is expected to result from the activities described. You may also find it easy to elicit statements from staff members about which aspects of the program are free to vary and which are not; this information too can shed light on the program's aims and rationale.

Record the program's rationale

Careful examination of the *rationale* underlying the program goes hand in hand with efforts to base the program on a clear and consistent plan. The rationale on which any program is based, sometimes called a process model, is simply a statement of *why* this program—a particular set of implemented materials, activities, and administrative arrangements—is expected to produce the desired outcomes. Sometimes the relationship between methods and goals is transparent; but other times, particularly with innovative programs, the credibility of the program requires that the staff explain and justify program methods and materials.

Example. A team of teachers from four high schools in a large metropolitan area planned a work-study program. The purpose of the program was to teach *careering savvy*. The teachers defined this as "knowledge about what it takes to be successful in one's chosen field of endeavor." The district assigned a consultant to the project, Anna Smith, whose job it was to help teachers iron out administrative details involved with coordinating student placement. Ms. Smith had also been told to serve in whatever formative evaluation capacity seemed necessary.

Having discovered that the teachers did not write a proposal for the program, she asked that they meet with her so that she could write a short document describing the program's major goals and outlining at least the skeleton of the program. At this meeting the teachers described the basic program. Students would choose from among a set of community-wide jobs made available at minimum pay by various professional and business firms. The students would work as office clerks, sales persons, receptionists; they might be called on to make deliveries and do odd jobs.

Instantly Ms. Smith saw that the program was without a clear rationale. "What makes you think," she asked, "that students will gain an understanding of the important skills involved in carrying on a career as a result of their taking on menial jobs?" The staff had to admit that the program as planned did not guarantee that students would learn about the duties of people in different careers or about prerequisite skills for success. Together with Ms. Smith they restructured the program as follows:

- They added an observation-and-conversation component to ensure that sponsoring professionals and business people would commit some time to describing their personal career histories and would allow the students to observe the course of their work day.
- Students would be required to keep journals and read about the career of interest.

The formative evaluator should see to it that the program rationale is suited to the conditions under which the program will be carried out. A mismatch may arise, for example, from staff insensitivity to time needed for the program to produce its effects. This could be reflected in too many objectives or objectives that are too ambitious for a project of moderate duration.

Lack of a clear program statement does not necessarily mean that goals, a program rationale, and plans for activities do not exist.[3] Producing a program statement is most often a matter of helping the staff and planners to coordinate their intentions and shape their ideas.

Production of the *written* statement provides a good opportunity for planners to describe concretely the program they envision. Because you have the job of *writing* an official statement *for the staff*, you will be able to ask difficult questions without implying any criticism. In describing goals and activities, and especially in exploring the logic of the connection between them, the program staff may encounter contradictions, uncertainties, and conflicts that you will have to handle with tact, patience, and persistence. Their sense of ease with you in your evaluator role will be reflected in the degree of candor with which they participate in these discussions. Interviews with program staff and first-hand observations might need to replace group discussions as your primary source of information if staff members find it too hard to articulate goals, strategies, and rationale in group settings.

Work on Agenda B can proceed concurrently with work on Agenda A. Since you will be meeting with the program staff to reach agreement on your relative roles, you might also use these meetings to clarify program goals, describe implementation plans, and work out the rationale.

The staff, finally, should keep in mind that during a formative evaluation, the program statement you produce is a *working document*. You, and they, will update it periodically, perhaps at the end of each reporting interval you have agreed upon. In the meantime, the existing document will be useful to people interested in the program. Besides guiding both the program as implemented and the evaluation, it can serve as the basis of reports and public relations documents.

---

[3] If, in fact, there is a total lack of consensus concerning what the program is about, you may find it necessary to do a retrospective needs assessment with the staff. Needs assessments result in lists of prioritized goals, determined by polling the wants of the educational constituency and determining how well these wants are being filled by the current program. Needs assessment is discussed in greater detail on page 8.

Agenda C: ~~Monitor Program Implementation and the Achievement of Program Objectives~~ collect information and analyze data

One of the distinctive features of formative evaluation is the continuous description and monitoring of the program as it develops, including measurement of the impact it is having on the attitudes and achievement of its target groups. The first two agendas focus on tentative agreements about the program's scope, rationale, planned activities, and goals. With Agenda C, you begin to investigate the match between the paper program, filled with intentions and plans, and the program in operation. The information you gather about the program for Agenda C can be used to:

- Pinpoint areas of program strength and weakness
- Refine and revise the program statement and, possibly, your evaluation plan
- Hypothesize about cause-effect relationships between program features and outcomes
- Draw conclusions about the relative effectiveness of program components where you have been able to use good evaluation design and credible measures

Agenda C is the phase of ~~formative~~ evaluation with the strongest research flavor. It can involve selecting samples; developing, trying out, selecting, administering, and scoring instruments; and analyzing and interpreting data. The component books of the *Program Evaluation Kit* will be most useful for this phase of the evaluation. In addition, if you will be conducting pilot studies, or if your evaluation can use a randomized control group, then you can refer to Chapter 5 the *Step-by-Step Guide For Conducting a Small Experiment* for more precise guidance.

In carrying out Agenda C, as with the others, you and the audience share information with a view toward producing a product. In this case, the product is an analysis, sometimes summarized in an interim report, of the program's implementation and its progress toward achieving its objectives. You tell the program staff at this point about the specifics of your sampling plan and site selection, and the measures you have chosen to purchase or construct in order to study features of program implementation or to monitor the attitudes or achievement of different subgroups. You might, in addition, describe the pilot tests or case studies you have chosen to pursue.

The first task of the *program staff* during Agenda C is to respond to your data gathering plan, suggesting adjustments to focus it more closely on what they most want to know. They should also confer with you in order to ensure that your measurements will not be too intrusive on program activities or personnel. Finally, they should share with you their perceptions of the credibility of the information you propose to collect.

Once you have reported results to the staff and planners from one round of data collection, the audience's job will be, quite naturally, to carefully examine what you have said and choose a course of action.

The degree to which *your own personal opinions* should guide your data collection and reporting is, incidentally, something to be negotiated with staff and planners. You could, on the one hand, take the stance of an impartial conduit for transporting the information the staff feels it needs. At another extreme, you could be highly opinionated, calling staff attention to what you feel are the program's most critical and problematic processes and outcomes. In the former situation, your report will convey to program planners the data you collected with the postscript, "Now *you* make the decision." If you plan to express opinions, then your reports will likely advocate a course of action, and the data collected will be planned with an eye toward providing evidence to support your case.

#Formative data collection plans

Ideally it would be nice if the formative evaluator could remain on-site with the program for extended periods of time, in the style of the participant-observer. Realistically, however, it is likely that budget, time, and possibly the geographical distribution of program sites, will make such vigilance impossible. You will have to rely on sampling, good rapport with the staff, and a well-designed measurement plan to give you an accurate picture of the program and its effects.

Your major source of first-hand information about the program will be your own informal observations and conversations with staff members while on site. Their descriptions of the program and explanations of what you see occurring should give you a good idea of how to design more formal data gathering instruments. Informal observations should also show where the program is going well and where it is failing, where a program component has been efficiently carried out, where it is partially implemented, and where it is not taking place.

In order to ensure that your informal impressions are representative and accurate, more formal data gathering will be necessary. For the purpose of formative evaluation, three approaches to collecting data about the program seem most useful:

- Periodic program monitoring
- Unit testing
- Pilot and feasibility studies

Your choice will be primarily determined by what you want to know.

Periodic program monitoring. The formative evaluator who wishes to check for proper program implementation throughout the evaluation selects a target set of characteristics which he then monitors periodically and at various sites. He also may select or construct achievement tests and attitude instruments to assess at these times the attainment of objectives of interest to the staff. The sites supplying formative information and the times at which this information is collected are often based on a sampling plan to ensure that the measurements made at intervals reflect the program as a whole.

2-13

BEST COPY

Example. Leonard Pierson, assistant to a district's Director of Research and Evaluation, was asked to serve as formative evaluator during the first year of a parent education program. The purpose of the program was to train parents of preschool children to tutor them at home in skills related to reading readiness: classification of objects, concept formation, basic math and counting, conversation and vocabulary. Federal funds had been provided for the training and to purchase home workbooks which were supplied free of charge. These workbooks sequenced and structured the home tutoring. They contained lessons, suggestions for enrichment activities, and short periodic assessment tests. The parent training centers were set up at six community agencies and schools throughout the district. Local teachers conducted evening classes to teach parents to use the workbooks daily with their children at home.

When the project director contacted Mr. Pierson, he was simply asked to give whatever formative evaluation help he could. Mr. Pierson, free to define his own role, decided to focus on four questions:

- Most importantly, do students learn the skills that are emphasized by the workbook?
- To what extent do parents actually work with their child daily?
- Do parents use techniques taught them in the training course, or do they develop their own?
- Are their own techniques more or less effective than those they have been trained to use?

In order to help answer these questions, Mr. Pierson designed two instruments and a monitoring system for administering them periodically:

- A general achievement test consisting of items sampled from the progress tests in the workbooks. The test will be administered every six weeks to a sample of participants' children. Presumably scores on this test should increase over time. A control group will also take the test every six weeks to account for learning due to sheer maturation.
- An observation instrument to be completed by the community member who visits the home to give the six weekly achievement tests. The instrument records the amount of progress made in the workbooks since the last visit, the nature of the teaching methods used by the parent, and the apparent appropriateness of the current lesson to the student's skills—that is, whether it seems to be too difficult. Observers will be trained to be particularly alert to changes in teaching style, recording both deviations and innovations.

The details of a periodic monitoring plan are usually agreed to by the evaluator and planners at the beginning of the formative evaluation and then vary little throughout the evaluator's collaboration with the program. The periodic program monitor submits interim reports at the conclusion of each data gathering phase. Like Table 3, these often focus on whether the program is on schedule.

A formative evaluator can use Table 3 to report to the program director and the staff at each location the results of monthly site visits. Each interim report could include an updated table accompanied by explanations of why ratings of "U," unsatisfactory implementation, have been assigned. The occasion at which measurements are made are determined by the passage of standard intervals--a month, a semester--or by logical transition periods in the program--such as the dates of completion of critical units. The evaluator might check time and again at the same sites or with the same people, or he could select a different representative sample to provide data at each occasion.

TABLE 3
Project Monitoring--Activities[4]

Objective 6: By February 29, 19YY, each participating school will implement, evaluate results, and make revisions in a program for the establishment of a positive climate for learning.

Winona School District
Wiley School

	19XX				19YY			
Activities for this objective	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Ap
6.1 Identify staff to participate		I	C					
6.2 Selected staff members review ideas, goals, and objectives		I	P	P	C			
6.3 Identify student needs		U	I	P	C			
6.4 Identify parent needs		U	I	P	C			
6.5 Identify staff needs		U	I	P	C			
6.6 Evaluate data collected in 6.3 - 6.5						I	U	C
6.7 Identify and prioritize specific outcome goals and objectives			I	U	P	P	C	
6.8 Identify existing policies, procedures, and laws dealing with positive school climate		U	I	P	P	C		

Evaluator's Periodic Progress Rating:
I = Activity Initiated P = Satisfactory Progress
C = Activity Completed U = Unsatisfactory Progress

Where the same measures are used repeatedly at the same sites, periodic monitoring resembles a time-series research design. This permits the evaluator to form a defensible interpretation of the program's role in bringing about the changes recorded in achievement and attitudes. Using a control group whose progress is also monitored further helps the evaluator to estimate how program students would be performing if there were no program.

Unit testing. An evaluator can focus on individual units of instruction or segments of the program that the staff has identified as particularly critical or problematic. In this case, monitoring of implementation will require *in-depth* scrutiny of the *particular* program component under study. Because the evaluator's task is to determine the value of specific program components, the implementation of these components will need to be described in as detailed a way as possible. Achievement tests, attitude instruments, and other outcome measures will have to be sensitive to the objectives that units of interest address. This could make it necessary for the evaluator to tailor-make a test, since general attitude and achievement tests will be unlikely to address the particular outcomes of interest. In some cases, the curriculum's own end-of-unit tests can be administered. If you use curriculum embedded tests, however, be careful that they are not so filled with the program's own format and content idiosyncracies that they sacrifice generalizability to other contexts or make the control group's performance look misleadingly bad. The occasions on which measurements are made for unit testing are determined by when important units occur during the course of the program. Sampling of *sites* and *participants* should be done where it is inadvisable or impractical to measure all students, but representative subgroups of students or classrooms can be measured or observed.

---

[4] This table has been adapted from a formative monitoring procedure developed by Marvin C. Alkin.

2-14

BEST COPY

Reports about the effectiveness of these critical program events should be delivered in time for modifications to be made in similar units, or in the same units in preparation for the next time a group of program participants encounters them.

If the teaching of units can be staggered at different sites so that not all students are being taught the same unit at the same time, then the results of unit tests can be used to make decisions about the best way to teach that unit at sites where it has not yet been introduced. Using a *control group* for unit testing gives quicker information about the relative effectiveness of different ways to implement the unit in question–more than one version can be tried out at the same time and their relative effectiveness assessed. Unit testing with a control group amounts to the same thing as a pilot or feasibility study.

Pilot and feasibility studies. These are usually undertaken because members of the program staff or its planners have in mind a particular set of issues that they need to settle or a hard decision to make. Pilot and feasibility studies are carefully conducted and usually experimental efforts to judge the relative quality of two or more ways to implement a particular program component. Pilot studies could be undertaken, for instance, to determine the most effective order in which to present information in a science discovery lab or the most beneficial time to switch students in a bilingual program to an all-English reading group. These studies require that different competing versions of a program component be installed at various sites. The evaluator first checks the degree to which each site carried out the program variation it was assigned; then, after giving the variation time to produce the results, he tests for their relative effectiveness. Like unit testing, feasibility studies demand measurement instruments that are sensitive to the outcomes that the program versions aim to produce. They usually demand random sampling since they use statistical tests to look for significant differences in the performance of groups experiencing different program variations. Pilot tests generally take place either before the program has begun, or *ad hoc* throughout the course of the evaluation whenever controversy or lack of information creates a need to try variations of the program.

Example 1. Dr. Schwartz, the university professor working as a formative evaluator for educational television station KDKC, overheard a conversation one day between two writers working on an episode for a series on cultural awareness. "*Poverty and Potatoes* is a silly name for an episode on Ireland," one writer was saying. "Well, maybe you can do better; but I say we need catchy titles–things that will make the kids want to watch," the other writer retorted. Dr. Schwartz offered to help the writers find such a title. He suggested that for each episode of the program, they write four or five possible titles. He would then construct and administer a questionnaire for students in order to find out from the target audience which title would most entice them to watch a television program.

Example 2. The Stone City school board voted in January to dismantle special classrooms for the educationally handicapped and return EH students to the regular classroom beginning in September. EH students would spend most of their time in the regular classroom, but would be "pulled out" each day to work with a special education teacher at a resource room in the school. The change would mean not only shifting students and altering the job roles of special education teachers; it would demand establishing resource rooms in schools which did not previously have them.

Faced with this large change of delivery of special education services, the district Director of Special Education suggested that some pilot work done during the present school year could prevent mistakes when mainstreaming took effect in the whole district in September. With board approval, she decided to phase mainstreaming into eight of the district's schools during the spring:

Phase I. Two schools which already had resource rooms would move EH students into the regular classrooms in March. The Director of Special Education would carefully observe and informally interview teachers, students, and special education teachers at these two schools to identify major problems involved in the transition. She would then work out an instructional package for teachers and parents that could be used to alleviate some of the problems and misunderstandings that could coincide with the organizational change.

Phase II. In April, three additional schools would be mainstreamed, using the training and counseling package developed during the first phase. Again, the effectiveness and smoothness of the transition to mainstreaming would be assessed, based on observations and interviews with regular teachers, special education teachers, students, and parents. The April sample would include one school which had not used resource rooms in the past. This would give the director a notion of the effect of mainstreaming in situations where it represents an even larger departure from regular practice. The training package would be revised based on feedback from teachers and parents with whom it was used.

Phase III. In May, three schools which had not previously had resource rooms would be converted to mainstreaming. The experience of the first two phases hopefully would make this transition a smooth one.

The Director of Special Education, having experienced several months of work in mainstreaming, could spend the summer preparing materials for parents and training teachers to anticipate September's reorganization.

Pilot and feasibility tests usually occur only when the evaluator offers to do them. Planners do not usually ask to have this sort of service performed for them. A feasibility study need not, as well, be based on achievement outcomes. A common question it might address is "Will people like Version X better than Version Y?"

Whatever plan you use for monitoring the program's effects, your efforts will allow the staff to make data-based judgments about whether program procedures are having an effect on participants. Besides the outcomes that planners hope to produce, you will need to look vigilantly for *unintended* outcomes–side-effects that can be ascribed to the program but which have not been mentioned by planners or listed in official documents. Although side-effects are generally thought of as negative, they could as easily be beneficial. You might discover, for example, potentially effective practices spontaneously implemented at a few sites that are worth exporting to others. Negative unintended effects are important to discover if the program is to be improved. They highlight areas that require added attention, modification, or even discarding.

Remember that when the data you collect suggest revisions in the program, you will have to amend the program

2-15

TABLE 4
Contrasts Between Reports for Formative and Summative Evaluation

	Formative Report	Summative Report
Purpose	Shows the results of monitoring the program's implementation or of pilot tests conducted during the course of the program's installation. Intended to help change something going on in the program that is not working as well as it might, or to expand a practice or special activity that shows promise.	*Documents* the program's implementation either at the *conclusion* of a developmental period or when it has had sufficient time to undergo refinement and work smoothly. Intended to put the program on record, to describe it as a finished work.
Tone	Informal	Usually formal
Form	Can be written or audiovisual; can be delivered to a group as a speech, or take the form of informal conversations with the project director or staff, etc.	Nearly always written, although some formal, verbal presentation might be made to supplement or explain the report's conclusions.
Length	Variable	Variable, but sufficiently *condensed* or *summarized* that it can be used to help planners or decision makers who have little time to spend reading at a highly detailed level.
Level of specificity	*High*, focusing on particular activities or materials used by particular people, or on what happened with particular students and at a certain place or point in time.	Usually more *moderate*, attempting to document general program characteristics common to many sites so that summary statements and general, overall decisions can be made.

statement as well. Program staff should take part in making these revisions, and consensus should be reached before any changes are recorded in the program's official description. New program statements may also suggest revisions or additions to your contracted evaluation activities.

Agenda D: Report and Confer With Planners and Staff

The reporting mode for formative evaluation varies with the situation. As is shown in Table 4, formative reports almost never look like the more technical ones submitted by summative evaluators. Most formative reporting takes place in conversations or discussions that the evaluator has with individuals or groups of program personnel. The form of *your* report will depend on:

- The reporting style that is most comfortable to you and the staff with whom you are working
- The extent to which *official* records are required
- Whether you will disseminate results only among program sites, or to interested outsiders and the general community as well
- How soon the information must reach its audience in order to be useful
- How the information will be used

Whether reports are oral or written is up to you. If additional planning or program modification will be based on the reports you give, then it is best to *discuss* program effects with the staff, perhaps at a problem solving meeting, so that remedies for problems can be debated and decisions made.

A *written* report provides a documentation of activities and findings to which the audience can continually refer and that can be used in program planning and revision. Written reports, however, take time to draft, polish, discuss, and revise. This is time that might be better spent collecting information and working on program development with the staff. In many cases, the best way to leave a written trace of the results of your formative findings will be to periodically revise the program statement you produced ~~as part of Agenda B.~~

Face-to-face meetings provide the staff and planners with a forum for discussion, clarification, and detailed elaboration of the evaluation's findings as well as the opportunity for making suggestions about upcoming evaluation activities. During conversational reports, you will be able to make requests for assistance in solving logistical problems or collecting data. Staff members might also want to express their problems or suggest new information needs.

A schedule for *interim reports* should be part of the formative evaluation contract. The program staff should indicate when or how frequently they wish to review the results of each evaluation activity. Interim reports on the progress of program development should contain results of completed evaluation components, a reiteration of tasks yet to be

2-16

BEST COPY

accomplished, and a full description and rationale for any changes in your responsibilities that may have to be negotiated.

Formative evaluation reports can include feedback of different sorts. At minimum, such a report will simply describe what the formative evaluator saw taking place—what the program looked like and what achievement or attitudes appeared to be the result. Depending upon his presumed expertise in such matters, the formative evaluator may also make suggestions about changes, point to places where the program is in particular need, and offer services to help remedy these problems. Your contributions along these lines will depend on your expertise and the contract you have worked out with the planners and staff.

If your evaluation service has focus~~s~~ed on pilot or feasibility studies, then your report will follow a more standard outline, although you may supplement the discussion of the results with recommendations for adaptation, adoption, or rejection of certain program components and perhaps outline further studies that are needed.

The tentative nature of instructional components in the formative stages of a program should be a recurring theme in your conversations with and reports to the staff. You will find that once ~~teachers and~~ staff are comfortable with program procedures, they will want to avoid making further changes in the program. The formative evaluator will have to make a conscious effort to keep the staff interested in looking at program materials and procedures with a view toward making them *yet more* appropriate, effective, and appealing for the students. Although evaluators will have the responsibility of overseeing the collection of information to support decisions about program revisions, the suggestions and active involvement of teachers in this decision-making process is crucial. Everyone on the program staff should understand why the formative evaluation is occurring and should be encouraged to take part.

For Further Reading

Alkin, M. C., Daillak, R., & White, P. *Using evaluations: does evaluation make a difference?* Sage Library of Social Research. Beverly Hills, CA: Sage Pubns., 1979.

Baker, E. L., & Saloutos, A. G. *Evaluating instructional programs.* Los Angeles, CA: Center for the Study of Evaluation, 1974.

Havelock, R. G. *Planning for innovation through dissemination and utilization of knowledge.* Center for Research on Utilization of Scientific Knowledge, Institute for Social Research, University of Michigan, January, 1971.

Lichfield, N., Kettle, P., & Whitbread, M. *Evaluation in the planning process.* Oxford: Pergamon Press, 1975.

Nash, N., & Culbertson, J. (Eds). *Linking processes in educational improvement.* Columbus, OH: University Council for Educational Administration, 1977.

Patton, M. Q. *Utilization-focused evaluation.* Beverly Hills, CA: Sage Publications, 1978.

Chapter 3

Step-by-Step Guides For Conducting an Evaluation

Chapter Two lists some of the myriad jobs of an evaluator. Keep in the general differences between the roles of a formative evaluator and a summative evaluator. The goal of a formative evaluator is to collect and share with planners and staff information that will led to improvement in a developing program. A summative evaluator has the responsibility for producing an accurate description of the program, complete with measures of its effects, that summarizes what has transpired during a particular time period. Results from a summative evaluation, usually compiled into a written report, canbe used for several purposes:

* To document for the funding agency that services promised by the program's planners have indeed been delivered
* To assure that a lasting record of the program remains on file
* To serve asa planning document for people who want to duplicate the program or adapt it to another setting.

The sometimes idiosyncratic nature of evaluations may make a step-by-step guide seem unnecessary, and in truth, there is no step-by-step way to perform tasks involved with the role.

Enough activities are common among evaluations, however, to permit a general outline of what needs to be accomplished. Chapter Two describes four Agendas to which an evaluator must attend to some degree. These agendas are:

* Agenda A: Set the Boundary of the Evaluation. That is, negotiate the scope of the data gathering activities in which you willengage, the aspects of the program on which you will

concentrate, and the responsibilities of your audience to cooperate in the collection of data and to use the information you supply.

Agenda B: Select Appropriate Evaluation Design and Measurements. In this case, you determine specifically what is to be measured, when, and how. The analysis to be employed is also planned at this stage.

Agenda C: Data Collection and Analysis. This includes administering all the planned data collection instruments to appropriate groups, compiling and analysing the data and selecting an appropriate reporting form.

Agenda D: Final Report. Given the original purpose of the evaluation, plan and execute a reporting strategy for the appropriate audiences.

As Chapter Two mentioned, many of the tasks falling within the scope of the different agendas will actually occur simultaneously or in an order other than that described by the guide. In general, however, there will be some logical order to how the evaluation unfolds. Consider the guide as a loose map of the activities you might perform.

If you are working as a formative evaluator for the first time in the setting, your best guidance might come from a conversation with someone who has evaluated the program before or who has served as formative evaluator in a similar setting. If formative evaluation presents a change in the evaluation role to which you are accustomed, then seek out someone who has done it before. Nothing beats advice from long experience.

Whenever possible, the step-by-step guides use checklists and worksheets to help you keep track of what you have decided and found out. Actually, the worksheets might be better called "guidesheets," since you will have to copy many of them onto your own paper rather than use the one in the book. Space simply does not permit the book to provide places to list large quantities of data.

As you use the guides, you will come upon references marked by the symbol 📖. These direct you to read sections of various *How To* books contained in the *Program Evaluation Kit*. At these junctures in the evaluation, it will be necessary for you to review a concept or follow a procedure outlined in one of these seven resource books:

- *How To Deal With Goals and Objectives*
- *How To Design a Program Evaluation*
- *How To Measure Program Implementation*
- *How To Measure Attitudes*
- *How To Measure Achievement*
- *How To Calculate Statistics*
- *How To Present an Evaluation Report.*

Al

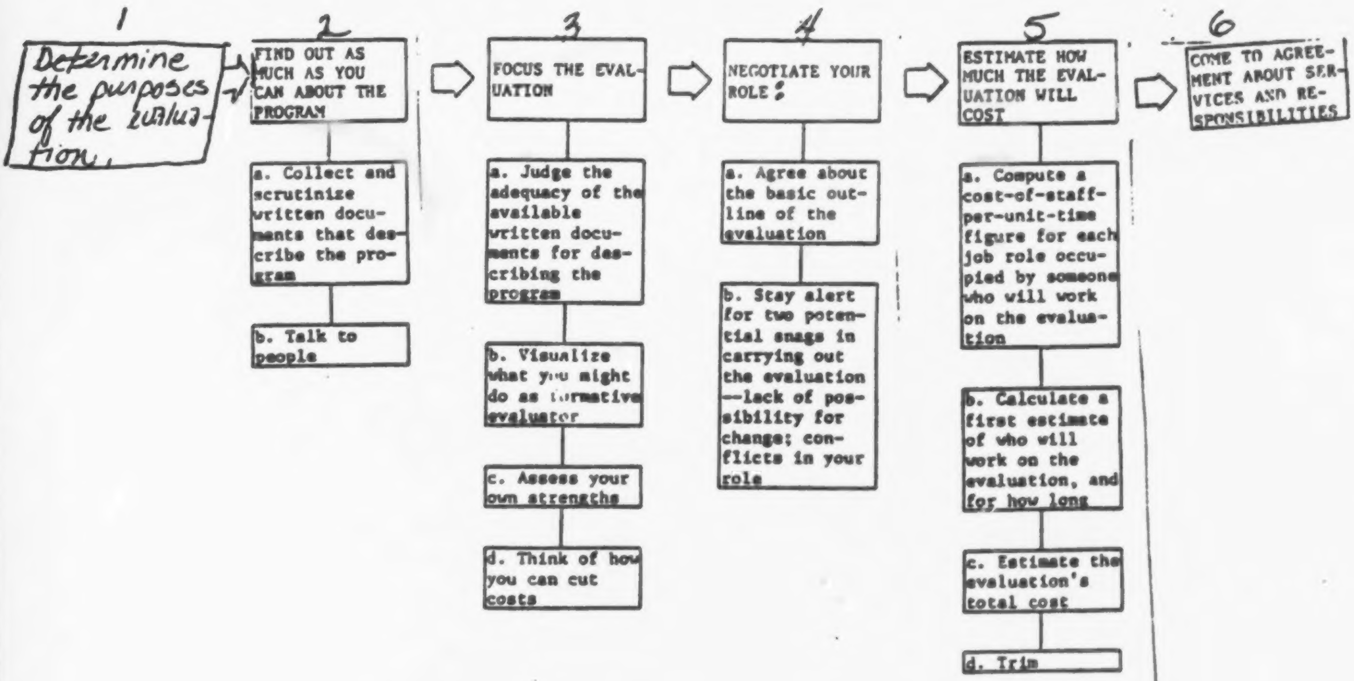
Agenda A.

Set the Boundaries For ~~of~~ the Evaluation

Instructions

Agenda A encompasses the evaluation planning period--from the time you accept the job of ~~formative~~ evaluator until you begin to actually carry out the assignments dictated by the role. Much of Agenda A amounts to gaining an understanding of the program and outlining the services you can perform, then negotiating them with the members of the staff who will use ~~formative~~ the information.

Agenda A's five steps and twelve substeps are outlined by this flowchart:



1
Determine the purposes of the evaluation.
2
FIND OUT AS MUCH AS YOU CAN ABOUT THE PROGRAM
a. Collect and scrutinize written documents that describe the program
b. Talk to people
3
FOCUS THE EVALUATION
a. Judge the adequacy of the available written documents for describing the program
b. Visualize what you might do as formative evaluator
c. Assess your own strengths
d. Think of how you can cut costs
4
NEGOTIATE YOUR ROLE
a. Agree about the basic outline of the evaluation
b. Stay alert for two potential snags in carrying out the evaluation --lack of possibility for change; conflicts in your role
5
ESTIMATE HOW MUCH THE EVALUATION WILL COST
a. Compute a cost-of-staff-per-unit-time figure for each job role occupied by someone who will work on the evaluation
b. Calculate a first estimate of who will work on the evaluation, and for how long
c. Estimate the evaluation's total cost
d. Trim
6
COME TO AGREEMENT ABOUT SERVICES AND RESPONSIBILITIES

3-4

BEST COPY

Step 1

Determine the purposes of the evaluation

Instructions

The job of the ~~summative~~ evaluator is to collect, digest, and report information about a program to satisfy the needs of one or more audiences. The audiences in turn might use the information for either of three purposes:

- To learn about the program or its components
- To satisfy themselves that the program they were promised did indeed occur and if not, what happened instead (particularly in a summative evaluation)
- To make decisions about continuing or discontinuing, expanding or ~~limiting~~ changing the program. ~~generally through giving or withholding funds~~

If ~~a~~ decisions hinge~~s~~ on your findings, your first job is to find out what the decisions are ~~is~~. Then you will have to ensure that you collect the appropriate information and report it to the correct audiences.

CHECK ☐ Begin your descriptions of the decisions to be made and your audience(s) by answering the following questions:

☐ What is the title of the program to be evaluated?________________

____

Throughout this chapter, this program will be referred to as Program X.

☐ What decisions will be based on the evaluation?

____

____

☐ Who wants to know about the program? That is, who is the evaluation's audience(s)?

- Teachers________________
  Report due________________
- Administrators________________
  Report due________________
- Counselors or department heads________________
  Report due________________
- District Personnel________________
  Report due________________
- School Board________________
  Report due________________
- Superintendent________________
  Report due________________
- State Department of Education________________
  Report due________________
- Federal Personnel________________
  Report due________________
- Parents________________
  Report due________________
- Community in general________________
  Report due________________
- Other--special interest groups, for instance
  ____
  Report due________________

NOTE Try not to serve too many audiences at once. To produce a credible ~~summative~~ evaluation, your position must allow you to be objective. ~~See pages 15 and 16 of this Handbook for elaboration of this critical admonition.~~

If you are conducting a summative evaluation, ~~A~~ask the people who constitute your primary audience this question:

☐ What would be done if Program X were to be found inadequate?

____

____

____

Here name another program or the old program, or indicate that they would have no program at all. What you enter in this blank is the alternative with which Program X should be compared. There could be many alternatives or competitors; but select the most likely alternative.

BEST COPY

3-5

A3

This most-likely-alternative-to-Program-X, its closest competitor, is referred to throughout this guide as Program C. Write it after the word "or" in the next sentence:

A choice must be made between continuing Program X or . . .

____

____

This is Program C.

If at all possible, set up or locate a control or comparison group which receives Program C.

~~Consult Chapter~~ [redacted] Program Evaluation for ideas [redacted] using [redacted] Chapter 1 describes evaluation designs, some of them fairly unorthodox, which might be useful for situations where control groups are difficult to set up. Using a control group greatly increases the interpretability of your information by providing a basis of comparison from which to judge the results that you obtain. Pages 24 to 32 of the same book describe different sets of control groups and the programs they might receive.

This approach may be changed, given new material in the Design and Qualitative Design books.

CHECK ☐

☐ Has one of the evaluation's audiences, such as a Federal or State funding agency, stated specific requirements for this evaluation? Are you required, for instance, to use particular tests, to measure attainment of particular outcomes, or to report on special forms? If so, summarize these evaluation requirements by quoting or referencing the documents that stipulate them.

____

____

____

____

____

____

CALENDAR

What is the absolute deadline for the earliest evaluation report? Record the earliest of the dates you listed when describing audiences.

~~The~~ An Evaluation Report must be ready by__________.

3-6

BEST COPY

Step 2

Find out as much as you can about the program(s) in question

Instructions

a. Scrutinize written documents that describe Program X, Program C, or both:

CHECK ✓ ☐

☐ A program proposal written for the funding agency

☐ The request for proposals (RFP) written by the sponsor or funding agency to which this program's proposal was a response

☐ Results of a needs assessment* whose findings the program is intended to address

☐ Written state or district guidelines about program processes and goals to which this program must conform

☐ The program's budget, particularly the part that mentions the evaluation

☐ A description of, or an organizational chart depicting, the administrative and staff roles played by various people in the program

☐ Curriculum guides for the materials which have been purchased for the program

☐ Past evaluations of this or similar programs

☐ Lists of goals and objectives which the staff or planners feel describe the program's aims

☐ Tests or surveys which the program planners feel could be used to measure the effects of the program, such as a district-wide year end assessment instrument

☐ Tests or surveys that were used by the program's formative evaluator, if there was one and you are conducting a summative evaluation.

☐ Memos, meeting minutes, newspaper articles—descriptions made by the staff or the planners of the program

☐ Descriptions of the program's history, or of the social context into which it has been designed to fit

☐ Articles in the education and evaluation literature that describe the effects of programs such as the one in question, its curricular materials, or its various subcomponents

☐ Other______________________________

Once you have discovered which materials are available, seek them out and copy them if possible.

Take notes in the margins. Write down, or dictate onto tape, comments about your general impression of the program, its context, and staff. This will get you started on writing your own description of the program. You may want to complete Step 3 concurrently with this general overview. Be alert, in particular, for the following details:

☐ The program's major general goals. List separately those that seem to be of highest priority to planners, the community, or the program's sponsors. Note where these priorities differ across audiences, since your report to each should reflect the priorities of each.

☐ Specifically stated objectives

☐ The philosophy or point of view of the program planners and sponsors, if these differ

☐ Examples of similar programs that planners intend to emulate

☐ Writers in the field ~~of education~~ whose point of view the program is intended to mirror

---

*A needs assessment is an announcement of educational needs, expressed in terms of the school curriculum and policies, by representatives of the school or district constituency.

BEST COPY

☐ The needs of the community or constituency which the program is intended to meet--whether these have been explictly stated or seem to implicitly underly the program

☐ Program implementation directives and requirements, described in the proposal, required by the sponsor, or both

☐ The amount of variation tolerated by the program from site to site, or even student to student

☐ The number and distribution of sites involved

☐ Canned or prepublished curricula to be used for the program

☐ Curriculum materials to be constructed in-house

☐ Plans which have been developed describing how the program looks in operation--times of day, scripts for lessons, etc.

☐ Administrative, decision-making, and teaching roles played by various people

☐ Staff responsibilities

☐ Descriptions of extra-instructional requirements placed on the program, such as the need to obtain parental permissions or to include teacher training or community outreach activities

☐ Student evaluation plans

☐ Plans for in-service training (especially for formative evaluations)

☐ Plans for curriculum development (for formative evaluations)

☐ Teacher evaluation plans

☐ Program evaluation plans

☐ Descriptions of program aspirations that have been stated as percents of students achieving certain objectives and/or deadlines by which particular objectives should be reached

☐ Timelines and deadlines for accomplishing particular implementation goals or reaching certain levels of student achievement (especially for formative evaluations)

b. Talk to people

in a formative evaluation, Once you have arrived at a set of initial impressions, check these--and your germinating evaluation plans--by seeking out people who can give you two kinds of information:

- Advice about how to go about collecting formative information for a program of this sort
- Answers to your questions about what the program is supposed to be and do--including which and how much modification can occur based on your findings

In any evaluation, check your description of the program against the impressions and aspirations of your audiences and the program's planners and staff. By all means, contact the people who will be in the best position to use the information you collect, your primary audience.

Try to think at this time of other people whose actions, opinions, and decisions will influence the success of the evaluation and the extent to which the information you collect will be useful and used. Make sure that you talk with each of these people, either at a group meeting or individually. Seek out in particular:

☐ Evaluators who have worked with this particular program or programs like it. They will have valuable advice to give about what information to collect, how, and from whom.

☐ School or district personnel not directly connected with the project, but whose cooperation will help you carry out the evaluation more efficiently or quickly. Negotiate access to the programs!

☐ Influential parents or community members whose support will help the evaluation go more smoothly

(continued)

☐ The project director(s)

☐ **Teachers, particularly those who seem to have most influence**

☐ **Program planners** and designers

☐ **Curriculum** consultants to the project

☐ **Members of advisory committees**

☐ Influential or particularly helpful students

☐ The people who wrote the proposal

If they are too busy to talk, send memos to key people. Describe the evaluation, what you would like them to do for you, and when.

In your meetings with these people, you should communicate two things:

- Who you are and why you are ~~formatively~~ evaluating the program
- The importance of your staying in contact with them throughout the course of the evaluation

They, in turn, should point out to you:

- Areas in which you have misunderstood the program's objectives or its description
- Parts of the program which will be alternatively emphasized or relatively disregarded during the term of the evaluation
- Their decision about the boundaries of the cooperation they will give you

Keep a list of the addresses and phone numbers of the people you have contacted with notations about the best time of day to call them or the times when it is easiest for them to attend meetings.

If possible, observe the program in operation or programs like it. Take a field trip in the company of program planners and staff. Have them point out the program's key components and major variations.

Take careful notes of everything you see and hear. Later you may find some of these valuable.

Step 3

Focus the evaluation

Instructions

a. Judge the adequacy of the available written documents for describing the program

DECISION

Make a note of your impressions of the quality and specificity of the program's written description. Answer these questions in particular:

- Are the written documents specific enough to give you a good picture of what will happen? Do they suggest which components you will evaluate and what they will look like?

  ☐ yes ☐ no ☐ uncertain

  ____

  ____

- Have program planners written a clear rationale describing why the particular activities, processes, materials, and administrative arrangements in the program will lead to the goals and objectives specified for the program?

  ☐ yes ☐ no ☐ uncertain

  ____

  ____

- Is the program that is planned, and/or the goals and objectives toward which it aims, consistent with the philosophy or point of view on which the program is based? Do you note misinterpretations or conflicting interpretations anywhere?

  ☐ yes ☐ no ☐ uncertain

  ____

  ____

If your answers to any of these questions is no or uncertain, then you will have to include in ~~your~~ Formative evaluation plans discussions with the planners and staff to persuade them to set down a clear statement of the program's goals and rationale.

b. Visualize what you might do as formative evaluator

DECISION

Base this exercise upon your impressions of the program:

- Which components appear to provide the key to whether it sinks or swims?________________

  ____

- Which components do the planners and staff most emphasize as being critically important?

  ____

  ____

- Which are likely to fail? Why?________________

  ____

- What might be missing from the program as planned that could turn out to be critical for its success?________________

  ____

- Where is the program too poorly planned to merit success?________________

  ____

- Which student outcomes will it probably be easiest to accomplish? Which will be most difficult?________________

  ____

- What effects might the program have that its planners have not anticipated?________________

  ____

While conducting this exercise by yourself, do not be afraid of being hard on the program. It is your job to foresee potential problems that the program's planners might overlook.

When you think about the service you can provide, you will, of course, need to consider two important things besides program characteristics and ~~outcomes. These are the budget~~ [illegible] ~~your own particular strengths and talents.~~

outcomes. These are the budget, which you will work out in Agenda B, and your own particular strengths and talents.

BEST COPY

3-10

c. Assess your own strengths

DECISION

You will best benefit the program in those areas where your visualization in Step 3b matches your expertise. You should "tune" the evaluation to build on your skills as:

- ☐ A researcher
- ☐ A group process leader or organizational facilitator
- ☐ A subject matter "expert"--perhaps a curriculum designer
- ☐ A former teacher in the relevant subject areas
- ☐ An administrator
- ☐ A facilitator for problem solving
- ☐ A counselor or therapist
- ☐ A linking agent (See page 27, Chapter 2)
- ☐ A good listener or speaker
- ☐ An effective writer
- ☐ A synthesizer or conceptualizer
- ☐ A disseminator of information, or public relations promoter
- ☐ Other________________________

d. Think of how you can cut costs

DECISION

Since the services you can envision providing probably exceed your budget, think of how you can cut costs. ~~See Step 2c, page 41.~~

[illegible]

Step 4

Negotiate your role

Instructions

Chapter 2 presented a general outline of the tasks that often fall within ~~the formative~~ An evaluator's role. You will have to work out your own job with your own audience. Meet again and confer with the people whose cooperation will be necessary--those whose decisions about the program carry most influence and who will cooperate when you gather information. You may, of course, also want to meet with other audiences.

a. Agree about the basic outline of the evaluation

CHECK ✓ ☐

☐ Agree about the program characteristics and outcomes that will be your major focus--regardless of the prominence given them in official program descriptions. Ask the planners and staff these questions:

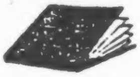
- Which characteristics of the program do you consider most important for accomplishing its objectives? Might you have implemented it in a different way than is currently planned? Would you be willing to undertake a planned variation study and try this other way?
- What components would you like the program to have which are currently not planned? Might we try some of these on a pilot basis?
- Are there particularly expensive, troublesome, controversial, or difficult-to-implement parts of the program that you might like to change or eliminate? Could we conduct some pilot or feasibility studies, altering these on a trial basis at some sites?
- Which achievements and attitudes are of highest priority?
- On which achievements and attitudes do you expect the program to have most direct and easily observed effect?
- Does the program have social or political objectives that should be monitored?

☐ Agree about the sites and people from whom you will collect information. Ask these questions:

- At which sites will the program be in operation? How geographically dispersed are they?
- How much does the program as implemented vary from site to site? Where can such variations be seen?
- Who are the important people to talk with and observe?
- When are the most critical times to see the program--occasions over its duration, and also hours during the day?
- At what points during the course of the program will it be best to measure student progress, staff attitudes, etc? Are there logical breaking points at, say, the completion of particular key units or semesters? Or does the program progress steadily, or each student individually, with no best time to measure?

3-12

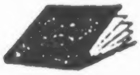
- Would it be better to monitor the program as a whole periodically, or should the effectiveness of various program subparts be singled out for scrutinizing, or both?



More detailed description of sampling plans is contained in How To Measure Program Implementation, pages 60 to 64. How To Design a Program Implementation, pages 33 to 45, describes decisions you might make about when to make measurements.

☐ Agree about the part the staff will play in collecting, sharing, and providing information. Explain to the staff that its cooperation will allow you to collect richer and more credible information about the program--with a clearer message about what needs to be done. Ask:

- Can records kept during the program as a matter of course be collected or copied to provide information for the evaluation?
- Can record-keeping systems be established to give me needed information?
- Will ~~your teachers and~~ staff be able to share ~~achievement~~ information with me or help with its collection? Are they willing to administer periodic tests to samples of ~~students?~~ individuals?
- Will staff members be willing and able to attend brief evaluation meetings or evaluation planning sessions?
- Will you be willing and able to take part in planned program variations or pilot tests? Will you be willing to respond to attitude surveys to determine the effectiveness of program components?
- Based on the information I collect, will you be willing to spend time on modifying the program through new instruction, lessons, organizational or staffing patterns?
- Are you willing to adopt a formative wait-and-see experimental attitude toward the program?



How To Measure Program Implementation describes ways to use records kept during the program to back up descriptions of its implementation. See pages 79 to 88.

☐ Agree about the extent to which you will be able to take a research stance toward the evaluation. Find out:

- Will it be possible to set up control groups with whom program progress can be compared?
- Will it be possible to establish a true control group design by randomly assigning participants to different variations of the program or to a no-program control group? Will it be possible to delay introducing the program at some sites?
- Can non-equivalent control groups be formed or located?
- Will I have a chance to make measurements prior to the program and/or often enough to set up a time series design?
- Will I be able to use a good design to underlie pilot tests or feasibility studies?
- Will I be able or required to conduct in-depth case studies at some sites?

In the case of formative evaluations, details about the use of designs ~~in formative evaluation~~ are discussed in How To Design a Program Evaluation. See in particular pages 14 to 19 and 46 to 51. Case studies are discussed in How To Measure Program Implementation, pages 31 and 32.

☐ In a Formative evaluation, agree about the extent to which you will need to provide other services. Ask the staff and planners these questions:

- Do you need consultative help that stretches my role beyond collecting formative data? Do you want my advice about program modifications, for instance? Or help with solving personnel problems?

BEST COPY

3-13

- Do you want me to serve to some degree as a linking agent? Should I, for instance, conduct literature reviews, seek consultation from similar projects, or search out services or additional people or funds to help the project?

- Should I take on a public relations role? Will you want me to serve as a spokesperson for the project? To give talks or write a newsletter, for example?

b. Stay alert for two potential snags in carrying out the evaluation

- Lack of possibilities for changing the program as the result of a formative evaluation.
- Conflicts in your own responsibilities to the program and the sponsor

The following questions are most pertinent in a formative evaluation.

☐ Look out for lack of commitment to change on the part of planners or staff. It will be fruitless to collect data to modify the program if someone will resist modifications. Before you begin scrutinizing the program or its various components, then, you should find out where funding requirements, staff opinion, or the political surround restrict altering the program. Ask in particular the following questions:

***See below

- On what are you most and least willing, or constrained, to spend additional money? What materials, personnel, or facilities?

- Where would you be most agreeable to cutbacks? Can you, for instance, remove personnel? If particular program components were found to be ineffective, would you eliminate them? Which books, materials, and other program components would you be willing to delete?

- Would you be willing to scrap the program as it currently looks and start over?

- How much administrative or staff reorganization will you tolerate? Can you change people's roles? Can you add to staff, say, by bringing in volunteers? Can you move people--teachers, even students--from location to location permanently or temporarily? Can you reassign students to different programs or groups?

- How much instructional and curricular change will you tolerate in the program beyond its current state? Would you be willing to delete, add, or alter the program's objectives? To what extent would you be willing to change books, materials, and other program components? Are you willing to rewrite lessons?

Include additional "what if..." questions that are more specific to the program at hand.

☐ Look out for conflicts in your own role. If your job requires that you report about the program to its sponsor or to the community at large, staff members are likely to be reluctant to share with you doubts and conjectures about the program. Since this will hamper your effectiveness, you will do best to explain to the planners and staff the following:

- ~~That you do not intend to write a summative report that judges and finds fault with the program.~~ Outline the form and some of the message that the report will contain.

  and/or

- That the planners and staff will have a chance to screen reports that you submit to the sponsor.

  and/or

- That you are willing to write a final report describing only those aspects of the program chosen by the staff.

  and/or

- That you are willing to swear confidentiality about the issues and activities that the evaluation addresses.

NOTE If you are in a hurry, and you think that you need to purchase instruments for the evaluation, then get started on this right away. ~~Consult Steps C3 and C4, and the relevant How To books, and order specimen sets as soon as possible.~~

** The following questions are most pertinent in a formative evaluation.

3-14

BEST COPY

9

~~exactly the same as [illegible]~~

Step ~~3~~ 5

Estimate the cost of the evaluation

Instructions

If your activities will be financed from the program budget, you will have to determine early the financial boundaries of the service you provide. The cost of an evaluation is difficult to predict accurately. This is unfortunate, since what you will be able to promise the staff and planners will be determined by what you feel you can afford to do.

Estimate costs by getting the easy ones out of the way first. Find out costs per unit for each of these "fixed" expenses:

☐ Postage and shipping (bulk rate, parcel post, etc.) ____________
☐ Photocopying and printing ____________
☐ Travel and transportation ____________
☐ Long-distance phone calls ____________
☐ Test and instrument purchase ____________
☐ Consultants ____________
☐ Mechanical test or questionnaire scoring ____________
☐ Data processing ____________

These fixed costs will come "off the top" each time you sketch out the budget accompanying an alternative method for evaluating the program.

The most difficult cost to estimate is the most important one: the price of person-hours required for your services and those of the staff you assemble for the evaluation. If you are inexperienced, try to emulate other people. Ask how other evaluators estimate costs and then do likewise.

Develop a rule-of-thumb that computes the cost of each type of evaluation staff member per unit time period, such as "It costs $4,500 for one senior evaluator, working full time, per month." This figure should summarize all expenses of the evaluation, excluding only overhead costs unique to a particular study--such as travel and data analysis.

The staff cost per unit figure should include:

Salary of a staff member for that time unit
+ Benefits
+ Office and equipment rental
+ Secretarial services
+ Photo copying and duplicating
+ Telephone
+ Utilities

→ This equals the total routine expenses of running your office for the time unit in question, divided by the number of full-time evaluators working there

Compute such a figure for each salary classification--Ph.D's, Masters' level staff, data gatherers, etc. Since the cost of each of these staff positions will differ, you can plan variously priced evaluations by juggling amounts of time to be spent on the evaluation by staff members in different salary brackets.

The tasks you promise to perform will in turn determine and be determined by the amount of time you can allot to the evaluation from different staff levels. An evaluation will cost more if it requires the attention of the most skilled and highly priced evaluators on your staff. This will be the case with studies requiring extensive planning and complicated analyses. Evaluations that use a simple design and routine data collection by graduate students or teachers will be correspondingly less costly.

To estimate the cost of your evaluation, try these steps:

a. Compute a cost-of-staff-per-unit-time figure for each job position occupied by someone who will work on the evaluation.

Depending on the amount of backup staff support entered into the equation, this figure could be as high as twice the gross salary earned by a person in that position.

b. Calculate a first estimate of which staff members' services will be required for the evaluation, and how long each will need to work.

BEST COPY

3-15

c. Estimate the evaluation's total cost.

CALENDAR

Refer to the proposed time span of the evaluation. Be sure to include fixed costs unique to the evaluation --travel, printing, long-distance phone calls, etc.--and your indirect or overhead costs, if any. Discuss this figure with the funding source, or compare it with the amount you know to be already earmarked for the evaluation.

d. Trim.

DECISION

Rather than visiting an entire population of program sites, for instance, visit a small sample of them, perhaps a third; send observers with checklists to a slightly larger sample, and perhaps send questionnaires to the whole group of sites to corroborate the findings from the visits and observations. See if one or more of the following strategies will reduce your requirment for expensive personnel time, or trim some of the fixed costs.

☐ Sampling

☐ Employing junior staff members for some of the design, data gathering, and report writing tasks

☐ Finding volunteer help, perhaps by persuading the staff that you can supply richer and more varied information or reach more sites if you have their cooperation

☐ Purchasing measures rather than designing your own

☐ Cutting planning time by building the evaluation on procedures that you, or people whose expertise you can easily tap, have used before

☐ Consolidating instruments and the times of their administration

☐ Planning to look at different sites with different degrees of thoroughness, concentrating your efforts on those factors of greater importance

☐ Using pencil-and-paper instruments that can be machine read and scored, where possible

☐ Relying more heavily on information that will be collected by others, such as state-administered tests, and records that are part of the program

"Instructions on How to Develop a Proposal" will go here.

Step ~~5~~ 6

Come to agreement about services and responsibilities

Instructions

~~An agreement outlining the duties of the formative evaluator and the program staff could conform to the following format:~~ The following agreement has been designed to outline the duties of a formative evaluator and program staff. Adjustments to the format could be made for a summative evaluation.

This agreement, made on __________, 19__ describes a tentative outline of the ~~formative~~ evaluation of the __________ project, funded by __________ for the academic year ______ to ______. The evaluation will take place from __________, 19__ to __________, 19__. The ~~formative~~ evaluator for this project is __________ assisted by __________ and __________.

Focus of the Evaluation

The program staff has communicated its intention that the ~~formative~~ evaluator monitor periodically the implementation of the following program characteristics and components across all sites:

____

____

Implementation of the following planned or natural program variations will be monitored as well:

____

____

The evaluator will monitor periodically progress in the achievement of these cognitive, attitudinal, and other outcomes:

____

____

The evaluator, in addition, will conduct feasibility and pilot studies to answer the following questions:

____

____

The evaluator will provide, as well, the following services to the staff and planners:

____

____

Data Collection Plans

Program Monitoring and Unit Testing

Data collection for ongoing formative monitoring of implementation and progress toward objectives will take place during the following periods: from __________ to __________; from __________ to __________; and from __________ to __________. These dates were chosen because ______________________________.

Interim reports, delivered to __________ and to __________, will be due on __________, 19__, __________, 19__, __________, 19__, and __________, 19__.

Approximately ______ program and ______ control sites for collection of implementation data will be chosen on a __________ (random/volunteer) basis. Of these, ______ will be studied intensively using a case study method; ______ will be examined by means of observation and interviews; and ______ will receive questionnaires or have records reviewed only. Staff members filling the following roles will be asked to cooperate:

____

Approximately ______ program and ______ control sites will take part in each assessment of progress toward program outcomes. These will be chosen on a __________ basis.

During each assessment period listed above, the following types of instruments will be administered to students and __________:

____

____

3-17

BEST COPY

Pilot and Feasibility Studies

Pilot and feasibility studies will be conducted at approximately ________ sites, chosen on a ________ basis. The purpose and probable duration of each study is outlined below:

____

____

Tentative completion dates for these studies are ________, 19__, ________, 19__, and ________, 19__, with reports delivered to ________ and ________ on ________, 19__, ________, 19__, and ________, 19__.

The following implementation, attitude, achievement, and other instruments will be constructed for the pilot studies: ________________

____

Staff Participation

Staff members have agreed to cooperate with and assist data collection during monitoring, unit testing, and pilot studies in the following ways:

____

____

Approximately ____ meetings will be needed to report and describe the evaluation's findings. These meetings, scheduled to occur a few days after submission of interim reports, will be attended by people filling the following roles:

____

The planners and staff have agreed that decisions such as the following might result from the formative evaluation: ________________

____

____

Budget

The evaluation as planned is anticipated to require the following expenditures:

Direct Salaries $ ________

Evaluation and Assistant Benefits $ ________

Other Direct Costs:
- Supplies and materials
- Travel
- Consultant services
- Equipment rental
- Communication
- Printing and duplicating
- Data processing
- Equipment purchase
- Facility rental $ ________

Total Direct Costs $ ________

Indirect Costs $ ________

TOTAL COSTS $ ________

Variance Clause

The staff and planners of the ________ program, and the evaluator, agree that the evaluation outlined here represents an approximation of the formative services to be delivered during the period ________, 19__ to ________, 19__.

Since both the program and the evaluation are likely to change, however, all parties agree that aspects of the evaluation can be negotiated.

The contract outlined here prescribes the evaluation's general outline only. If you plan to describe either the program or the evaluation in greater detail, then include tables such as **Tables 2 and 3 in Chapter 2, pages 28 and 31.**

3-18

BEST COPY

AGENDA B

Select Evaluation Design
and Appropriate Measures

Decide what to measure or assess	Select design or monitoring system	Plan analysis for each instrument	Choose Sampling strategy

(Many of the steps in Agenda B are still to be combined. It will be more efficient after I receive the revised Implementation book. I assume the Design book will not have changed substantially.)

3-19

~~**Step 2**~~

Set up the evaluation designs

Instructions

In Phase B you selected instruments with which to carry out measurements and you chose an evaluation design to determine when--and to whom--they would be administered. The purpose of this step is to help you ensure that the design is carried out.

Issues of design and random assignment are treated in depth in How To Design a Program Evaluation. ~~In this book you will also find step-by-step directions for setting up any of the six designs you have chosen to use.~~

The three checklists which follow are intended to help you keep track of the implementation of the design you have chosen. Set up the checklist that is relevant to your particular design. ~~Use it to keep track of important information and to check the completion of activities essential to the design.~~

Checklist for a Control Group Design With Pretest--Designs 1, 2, and 3

1. Name the person responsible for setting up the design__________

If the design uses a true control group:

2. Will there be blocking? ☐ yes ☐ no

(See How To Design a Program Evaluation, pages 149 and 150.)

3. If yes, based upon what?
☐ ability ☐ sex
☐ achievement ☐ other__________

4. Has randomization been completed?
☐ yes ☐ no Date__________

If the design uses a non-equivalent control group:

5. Name this group__________

6. List the major differences between the program and comparison groups--for example, sex, SES, ability, time of day of class, geographical location, age:
____
____
____

7. Has contact been made to secure the cooperation of the comparison group? ☐ yes
Date__________

8. Agreement received from (Ms./Mr.)__________

9. Agreement was in the form of (letter/memo/personal conversation/etc.)__________

10. Confirmatory letter or memo sent? ☐ yes
Date__________

11. Is there a list of students receiving the comparison program? ☑ yes ☐ no
Where is it?__________

In either case:

12. Name of pretest__________

13. Pretest completed? ☐ yes Date__________

14. Teachers (or other program implementors) warned:
☐ To avoid confounds? Memo sent or meeting held (date)__________
☐ To avoid contamination? Memo sent or meeting held (date)__________

(See How To Design a Program Evaluation, page 60.)

15. List of possible confounds and contaminations
____
____
____

16. Check made that both programs will span the same time period? ☐ Date__________

17. Posttest given? ☐ Date__________

3-20

BEST COPY

Checklist for a Time Series Design With Optional Non-Equivalent Control Group --Designs 4 and 5

1. Name of person responsible for setting up and maintaining design________
2. Names of instruments to be administered and readministered________

   ____
3. Equivalent form of instruments to be:

   ☐ Made in-house? ☐ Purchased?
4. Number of repeated measurements to be made per instrument________
5. Dates of planned measurements:

   ☐ 1st________
   ☐ 2nd________
   ☐ 3rd________
   ☐ 4th________
   ☐ 5th________
   ☐ 6th________
   Additional:
   ☐ ________

If the design uses a control group:

6. Name of control group________
7. List of major differences between the program group and the control group--for example, sex, SES, ability, geographical location, age

   ____

   ____
8. Contact made to secure cooperation of comparison group? ☐ Date________
9. Agreement received from (Ms./Mr.)________
10. Confirmatory letter or memo sent? ☐

    Date________
11. List of possible contaminations

    ____

    ____

    ____

Checklist for Pre-Post Design With Informal Comparisons--Design 6

1. Name of person responsible for setting up design________
2. Comparison to be made between obtained posttest results and pretest results? ☐
   - Name(s) of instrument(s) to be used

     ____

     ____
   - Equivalent forms of instruments to be:

     ☐ made ☐ purchased
   - List of students receiving Form A on pretest and Form B on posttest________

     ____

     ____
   - List of students receiving Form B on pretest and Form A on posttest________

     ____

     ____
   - Dates of planned measurements:

     Pretest________ Completed? ☐
     Posttest________ Completed? ☐
3. Comparison to be made via standardized tests? ☐
   - Name of standardized test(s)________

     ____
   - Test given? ☐ Date________
   - Scoring and ranking of program students completed? ☐ Date________
4. Comparison to be made between obtained results and results described in curriculum materials? ☐
   - Name of curriculum materials________

     ____
   - Unit test results collected and filed? ☐
   - Unit test results from program graphed or otherwise compared to norm group? ☐
5. Comparison to be made between results from a previous year and the results of the program group? ☐
   - Which results from last year will be used--for example, grades, district-wide tests?

     ____
   - Last year's results tabulated and graphed? ☐
   - List made of possible differences between this and last year's (or last time's) group that might differentially affect results? ☐

     ____

     ____

     ____
   - Program X's results collected? ☐
   - Program X's results scored and graphed, or otherwise compared, with last year's? ☐

3-21

6. Comparison to be made between obtained results and prespecified criteria about attainment of program objectives? ☐

   - Whose criteria are these--for example, teachers, district, curriculum developers?
     ____

   - State the criteria to be met______________
     ____
     ____

   - Objectives-based test results collected and filed? ☐

   - Objectives-based test results graphed, or otherwise compared, with criterion? ☐

CHECK ☐ If you have chosen to administer instruments at only a sample of program sites or to a sample of respondents, then use the following table to keep track of the proper implementation of your sampling plan.

Sampling Plan Checklist

1. The sample will ensure adequate representation to different types of:
   - ☐ Sites--what kinds?______________
   - ☐ Time periods--which ones?______________
   - ☐ Program units--which ones?______________
   - ☐ Program roles--which ones?______________
   - ☐ Student or staff characteristics--name them
     ____
   - ☐ Other______________________________

2. The sampling plan comprises a matrix or cube with ______ cells (see How To Measure Program Implementation, pages 60 to 65)

3. How many cases will be sampled from each cell? ____________ (see How To Design a Program Evaluation, pages 157-161, for suggestions about selecting random samples)

4. Cases selected? ☐

5. For each time selected:
   - Have instruments been administered? ☐
     Comments______________________________
   - What deviations from the sampling plan have occurred?______________________________
     ____

3-22

AGENDA C

Collect and Analyze
Data According to
Evaluation Design

Put Sampling Plan into Effect	Administer instruments--observe, score, record	Compile& Reduce data	Analyze data

(Some of the steps in Agenda C are to be revised further when I have received the Implementation and the Reporting books)

~~Step~~ 3

Administer instruments, score them, and record data

Instructions

a. Once you have decided which instruments to use, begin acquiring them at once. Have no illusions--ordering and constructing instruments will take a long time, possibly months.

If you intend to buy instruments, use the form letters in the various How To books for ordering them. Check the list of test publishers in the How To books for sources of published tests.

If you plan to construct your own instruments, write a memo to those in charge of producing them, leaving no doubt about who is responsible and deadlines for their completion.

Instruments made in-house must be tried out, debugged, and evaluated for technical quality. To aid the process, the Kit's measurement books discuss reliability and validity as they apply to the three primary measurement concerns treated in the books. **See Achievement, Chapter 5, Attitudes,** Chapter 11, and **Implementation, Chapter 7. A little run-through with a few students or aides might mean** the difference between a mediocre instrument and a really excellent one.

Keep tabs on instrument orders. If you have not received them within two weeks of the deadline, prod the publisher or your in-house developer.

Once each instrument is completed or received, plan how it will be scored and recorded:

b. Score instruments as the results come in

If the instrument has a selected response format--for instance, multiple-choice, true-false, Likert-scale--make sure you have a scoring key or template.

If it has an open-ended format, make sure you have a set of correctness criteria for scoring, or a way of categorizing and coding questionnaire or interview responses.

See How To Measure Program Implementation, pages 71-73 and How To Measure Attitudes, pages 106 and 107, and 170 and 171. These sections contain information about scoring or coding open-response items, essays, and reports. If the test is to be scored elsewhere by a state or district office or by an agency with whom you have a contract for testing and scoring, and you are to receive a print-out of the results, decide whether you wish to score sections of it for your own purposes. In some cases, achievement of objectives can be measured via partial scoring of a standardized test.

~~See How To Deal With Goals and Objectives, pages 43 to [illegible],~~ and **How To Measure Achievement, pages 36 to 39 for a description of a technique** for doing this.

c. Record results per measure onto a data summary sheet

Once you know what the scores from your instruments will look like, decide whether you want results for each examinee, mean results for each class, or percentage results for each item. Then, when each instrument has been administered, score the instruments as soon as possible.

Once scoring is completed, consult the appropriate How To books for suggestions about formatting and filling out data summary sheets. **See Attitudes, pages 159 to 166; Implementation, pages 67 to 71; and Achievement, pages 117 to 120.**

Construct separate data summary sheets for Program X people and the comparison group so that it is impossible to get them confused. Then delegate the scoring and recording tasks.

3-24

BEST COPY

CHECK ✓ ☐

A table like the following should help you keep track of instrument development, administration, scoring, and data recording.

Instrument	Completion/ Receipt Deadline	Administration Deadlines				Scoring Deadline	√?	Recording Deadline	√?
		pre	√?	post	√?				

[redacted]

Analyze data with an eye toward implications for policy and for program improvement

Instructions

If you are periodically monitoring the program--and particularly if there is a control group--then you have collected a battery of general measures that can be analyzed using fairly standard statistical methods. Consider whether you will:

☐ Graph results from the various instruments

☐ Perform tests of the statistical significance of differences in performance among groups or from a single group's pretest and posttest

☐ Calculate correlations to look for relationships

☐ Compute indices of inter-rater reliability

How To Measure Achievement discusses using test results for statistical analysis on pages 125 to 145. How To Measure Attitudes describes attitude test scores used for calculating statistics on pages 170 to 177. See, as well, How To Measure Program Implementation, pages 67 to 77. Problems of calculating inter-rater reliability are discussed in all three books. Specific statistical analyses are discussed in How To Calculate Statistics.

All of the Kit's How To books contain suggestions for building graphs and tables to summarize results. For each instrument you use, see the relevant How To book. Consult, as well, Chapter 4 of How To Present an Evaluation Report.

When each graph and statistical test is completed, examine it carefully and write a one-or-two sentence description that summarizes your conclusions from reading the graph and noting the results of the analysis.

Save the graphs and summary sentences that seem to you to give the clearest picture of the program's impact. These can be used as a basis for the Results section of your report.

NOTE Remember that in addition to describing program implementation and the progress in development of skills and attitudes of various participants, you may also need to note whether the program is keeping pace with the time schedule that has been mapped out.

If you have focused data collection on specific program units, or if you are conducting pilot tests, then in addition to performing statistical analyses, consider whether the program has achieved each of the objectives in question. In particular, examine these things:

- Student achievement
- Participants' attitudes about the program component in question
- The component's implementation

Below you will find four cases describing results you might obtain in a formative evaluation, with suggestions about what to do about each: Determinations of good, poor, and adequate performance should be based on the performance standards set in Step 6.

Case 1

- Achievement test results: good
- Program implementation: adequate
- Attitude results: poor

What to do? Check the technical quality of the instrument (see How To Measure Attitudes, pages 131 to 151). Find out what is causing bad morale:

☐ Is the program too easy? Pretest students for upcoming program units to see if they have already mastered some of the objectives.

☐ Is the program too difficult? If this complaint is widespread, try to alleviate the pressure of the work.

☐ Is this part of the program dull? The response to this depends on the students and subject matter. Try to find motivators for the students, or help teachers to invent ways to make instruction more appealing and relevant. If minor changes offer no promise, the staff is convinced of the importance of program objectives, and the rest of the program seems more interesting, then don't revise.

3-26

BEST COPY

Case 2

- Achievement test results: good
- Program implementation: poor
- Attitude results: good

What to do? First ask:

☐ Did the achievement test and the program component address the same objectives? *If not*, there's your answer! *If so*, check the technical quality of the implementation measure. See *How To Measure Program Implementation*, pages 129 to 138.

Then ask:

☐ What happened in the program instead of what was planned? Make sure that students did not learn from the mistakes they made while struggling through poor instruction. If possible, suggest that the instruction that *did* occur become officially part of the program.

Case 3

- Achievement test results: poor
- Program implementation: good
- Attitude results: good

What to do? First ask:

☐ Did students misinterpret test items in some way, and if so, how?

Then assure yourself that the objective underlying the test matches the objective underlying the instruction. If so, examine the technical quality of the achievement test. See *How To Measure Achievement*, pages 89 to 115.

Then ask:

☐ Was student performance during program implementation good? If so, check whether the *amount of practice* given to students was sufficient to allow them to master the objective.

☐ Was student performance on program tasks poor? If so, explore whether sufficient time was given for practice *and* whether students lacked *prerequisite* skills necessary to learn the material. You may need to give diagnostic tests to locate students' skill deficiencies. Check to see whether the instruction itself was difficult or confusing. Did students understand what was expected of them?

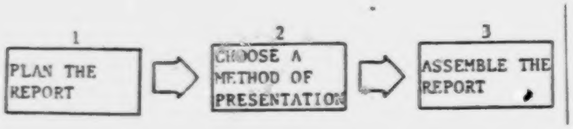
Case 4

More than two of the indicators show unsatisfactory results. In any of these cases, you should investigate the cause of the problem and revise as necessary.

AGENDA D

Report and Confer with Planners and Staff

The key to an effective evaluation is good communication. Especially in the case of a formative evaluation, information about where the program is or is not working needs to be timely and clearly presented so that appropriate changes may be made. Summative reports must also be timely and carefully prepared if they are to have an impact on policy decisions.


1
PLAN THE REPORT
2
CHOOSE A METHOD OF PRESENTATION
3
ASSEMBLE THE REPORT

3-28

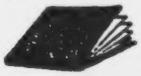
Step 1

~~Decide what you want to say~~ Plan the Report

Instructions

You want to get the information across quickly and succinctly. Therefore think about each instrument you have administered and:

☐ Make a graph or table summarizing the major quantitative findings you want to report. How To Calculate Statistics, pages 18 to 25, describes how to graph test scores.



☐ ~~Check How To Present an Evaluation Report, Chapters 1 and 2~~, for suggestions about organizing your message and an outline of an evaluation report. Look over the outline and decide which of the topics apply to your report. If you will need to describe program implementation, look at the report ~~outline in Chapter 2 of How To Measure Program Implementation.~~

☐ Write a quick general outline of what you plan to ~~discuss.~~ report.

If you are submitting an interim report for a program that is being assembled from scratch, you should include in Section II of the report a few paragraphs dealing with progress in program design. They might be entitled, for instance, Materials Production or Staff Development. The paragraphs should address these questions:

☐ Has research been conducted to determine the sort of curriculum that is appropriate to the program? Who conducted this research? How useful has it been?

☐ What materials development has been promised for the program? For which objectives? For which sites? What student materials? Any teacher manuals? Any teacher training materials? Any audiovisuals? Has the staff promised to expand or revise something previously existing? Did they submit in the proposal an outline, plan, or prototype of the promised materials? Are the materials being produced in accordance with this? Have there been changes? Has the staff decided to not develop something they promised? Why? How is development of these particular materials progressing? Is it on schedule? Behind? Why? Does the staff plan to catch up by year's end, or is this unnecessary because they are well ahead of student progress? How much of the intended materials development will be completed by the end of the evaluation?

☐ What staff development and training have been provided to ensure that planners, teachers, etc., are equal to the tasks of both designing and implementing a new program?

☐ What plans for staff member participation in materials development are contained in the proposal? Is this an accurate description of what has occurred so far?

☐ What staff-community interchanges to gather help with planning were mentioned in the proposal? What staff meetings--within the project or with staff members outside it--were planned? Did these occur? What were their purposes and outcomes?

3-24

Step 2

Choose a method of presentation

Instructions

If the manner of reporting was not negotiated during Agenda A, decide whether your report to each audience will be oral or written, formal or informal.

Chapter 3 of How To Present an Evaluation Report lists a set of pointers to help you organize what you intend to say and decide how to say it.

Step 3

Assemble the report

Instructions

Follow the outline described in Chapter 2 of How To Present an Evaluation Report, Section III. The report should include:

- A description of why you undertook the evaluation
- Who the decision-makers were
- The kinds of ~~formative~~ questions you intended to ask, the evaluation designs you used, if any, and the instruments you used to measure implementation, achievement, and attitudes
- The ~~best formal~~ data collection methods which you used

If you have found instruments which were particularly useful, or sensitive to detecting the implementation or effects of the particular program, put them in an appendix.

A formative ~~Your~~ report should conclude, importantly, with suggestions to the summative evaluator, if indeed a summative evaluation of this particular program will be conducted.

A worksheet like the one below will help you to record your decisions about reporting and to keep track of the progress of your report.

Final Report Preparation Worksheet

1. List the audiences to receive each report, date reports are due, and type of report to be given to each audience. Some reports may be suitable for more than one audience.

Audience	Date report due

2. How many different reports will you have to prepare? ______________________

3. For each different report you submit, complete this section:

Report #1 Audience(s)______________________

BEST COPY

D4

Checklist for Preparing Evaluation Report:

- Report will be: ☐ formal ☐ informal ☐ oral ☐ written
- Deadline for finished draft ____________
  Completed? ☐
- Deadline for finished audio-visuals, if any ____________
  Completed? ☐
- Deadline for finished tables and graphs ____________
  Completed? ☐
- Names of proofreaders of final draft, audio-visuals, or tables
  ____
  Contacted and agreement made? ☐
  ____
  Contacted and agreement made? ☐
  ____
  Contacted and agreement made? ☐
- Date agreed upon as deadline for getting drafts to proofreaders. These are absolute deadlines for completing drafts:
  ____________ Draft sent? ☐
  ____________ Draft sent? ☐
  ____________ Draft sent? ☐
- Dates drafts must be received in order to revise in time for final report deadlines:
  ________ Proofread draft received? ☐
  ________ Proofread draft received? ☐
  ________ Proofread draft received? ☐

This is the end of the Step-by-Step Guide for Conducting a ~~S[illegible]~~ Evaluation. By now evaluation is a familiar topic to you and, hopefully, a growing interest. This guide is designed to be used again and again. Perhaps you will want to use it in the future, each time trying a more elaborate design and more sophisticated measures. Evaluation is a new field. Be assured that people evaluating programs--yourself included--are breaking new ground.

BEST COPY

3-38

Chapter 5

4

Step-by-Step Guide For Conducting a Small Experiment

The self-contained guide which comprises this chapter will be useful if you need a quick but powerful pilot test—or a whole evaluation—of *a definable short-term program or program component.* The guide provides start-to-finish instructions and an appendix containing a sample evaluation report. This step-by-step guide is particularly appropriate for evaluators who wish to assess the effectiveness of *specific materials and/or activities* aimed toward accomplishing a *few specific objectives.*

If a major purpose of the program you are evaluating is to produce achievement results, this guide outlines *an ideal way to find out how good these results are: conduct an experiment.* For a period of days, weeks, or months, give students the program or program component you wish to evaluate while an equivalent group, the *control* group, does not receive it. Then at the end of the period, test both groups. This step-by-step guide shows you how to conduct such an evaluation.

Whenever possible, the step-by-step guide uses checklists and worksheets to help you keep track of what you have decided and found out. Actually, the worksheets might be better called "guidesheets," since you will have to copy many of them onto your own paper rather than use the one in the book. Space simply does not permit the book to provide places to list large quantities of data.

As you use the guide, you will come upon references marked by the symbol ◆. These direct you to read sections of various *How To* books contained in the *Program Evaluation Kit.* At these junctures in the evaluation, it will be necessary for you to review a concept or follow a procedure outlined in one of the Kit's seven resource books:

- *How To Deal With Goals and Objectives*
- *How To Design a Program Evaluation*
- *How To Measure Program Implementation*
- *How To Measure Attitudes*
- *How To Measure Achievement*
- *How To Calculate Statistics*
- *How To Present an Evaluation Report*

Should You Be Using This Step-By-Step Guide?

The appropriateness of this guide depends on whether or not you will be able to set up certain *preconditions* to make the evaluation possible. Check each of the preconditions listed in Step 1. If you can arrange to meet *all* of them, then you can use the evaluation strategy presented in this guide. As you assess the preconditions, you will be taking the first step in planning the evaluation. This step-by-step guide lists 13 steps in all. A flow chart showing relationships among these steps appears in Figure 5. You may wish to check off the steps as they are accomplished.

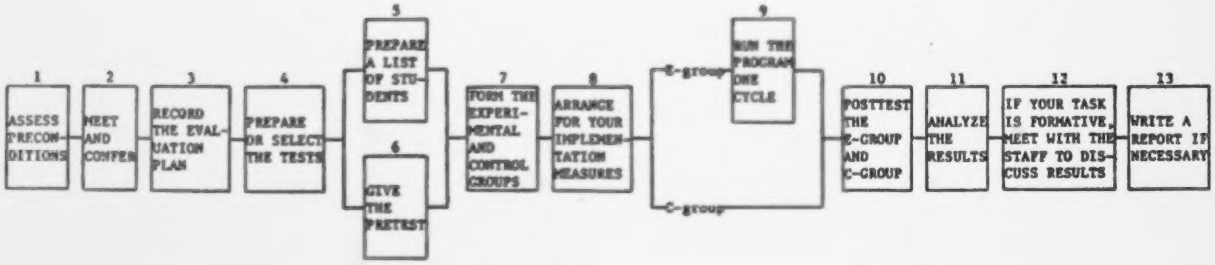

1
ASSESS PRECON-DITIONS
2
MEET AND CONFER
3
RECORD THE EVAL-UATION PLAN
4
PREPARE OR SELECT THE TESTS
5
PREPARE A LIST OF STU-DENTS
6
GIVE THE PRETEST
7
FORM THE EXPERI-MENTAL AND CONTROL GROUPS
8
ARRANGE FOR YOUR IMPLEMEN-TATION MEASURES
E-group
9
RUN THE PROGRAM ONE CYCLE
C-group
10
POSTTEST THE E-GROUP AND C-GROUP
11
ANALYZE THE RESULTS
12
IF YOUR TASK IS FORMATIVE, MEET WITH THE STAFF TO DIS-CUSS RESULTS
13
WRITE A REPORT IF NECESSARY

Figure 5. The steps for accomplishing a small experiment, listed in this guide

4-1

Step 1

Assess Preconditions

Instructions

CHECK ✓ ☐ Put a check in each box if the precondition can be met. For the first three preconditions, there are some decisions to be recorded on the lines provided. Record these decisions in pencil since you may change them later. This step-by-step guide will be useful to you only if you can meet all five preconditions.

☐ PRECONDITION 1. An outcome measure will be available.

A test can be made or selected to measure what students are supposed to learn from the program. Write down what the outcome measure(s) will probably be:

____

____

____

☐ PRECONDITION 2. A sample of cases* can be defined.

You can list at least 12, say, students for whom this program would be suitable and for whom, therefore, the outcome measure is an appropriate test of what they learned in the program. Write down the criteria that will be used to select students for the sample:

____

____

____

☐ PRECONDITION 3. A time period--a cycle--can be identified.

You can identify a time period which is of a duration appropriate to teach the skills the outcome measure taps. Call this period of time one cycle of the program. Write down what length of time one cycle of the program will probably last:

____

____

____

☐ PRECONDITION 4. An experimental group and a control group can be set up.

For one cycle at least, one group of students in the sample will get the program and another will not. If the program can run through several cycles, this does not mean that some students will never get the program, just that they must wait their turn. In this way, no students are left out--a concern which sometimes makes people unwilling to run an experiment.

☐ PRECONDITION 5. Students who are to get the program can be randomly selected.

The students who are to get the program during the experimental cycle will be randomly selected from the sample.

If each of the five preconditions listed above can be met, then you will be able to run a true experiment. This is the best test you can make of the effectiveness of the program or program component for producing measurable results.

---

*A case is an entity producing a score on the outcome measure. In educational programs, the cases of interest are nearly always students--though they could be classrooms, school districts, or particular groups of people. The word student is used throughout the guide. If the cases in your situation are different, just substitute your own term.

Step 2

Meet and Confer

Instructions

This step helps you work out a number of practical details that must be settled before you can complete your plans for the pilot test or evaluation.

CHECK ✓ ☐

You will need to meet and confer with the people whose cooperation you need and, possibly, with members of other evaluation audiences. You will need to reach agreement with them about:

- ☐ How the study should be run
- ☐ How to identify students for the program
- ☐ What program the control group should receive
- ☐ The appropriate outcome measure
- ☐ Whether to use additional measures
- ☐ What procedures will be used to measure implementation
- ☐ To whom results will be reported--and how

How Should the Study Be Run?

DECISION

In particular, are students to receive the program in addition to regular instruction or instead of regular instruction? If the program is to be used in addition to regular instruction, students will have to be pulled out for the program sometime other than the regular instruction period. A means of scheduling will need to be agreed upon.

How Should Students Be Identified for the Sample?

DECISION

It might be that the sample will simply be all the students in a certain class or classes. On the other hand, perhaps the program is intended only for students who have a certain need or meet some criterion. In this case, you will need to agree upon clear selection criteria. If the program is remedial, selection might be based on low scores on a pretest, or you might use teacher nominations. Test scores for selection are preferred if the outcome measure is to be a test. The problem with basing selection on an existing set of test scores is that they might be incomplete; scores might be missing for some students. You could use the outcome measure as a selection pretest.

How To Design a Program Evaluation, pages 35 and 36, discusses selection tests. See also How To Measure Achievement, pages 124 and 125.

How many students will you need? The more the better, but certainly you should avoid ending up with fewer than six pairs of students, a total of 12. If during the program cycle, one student in a pair is absent too often or fails to take the posttest, the pair will have to be dropped from the analysis. The longer the cycle, the more likely it is that you will lose pairs in this way. Bearing this in mind, be sure to select a large enough sample. If it looks as if the sample will be too small--perhaps because the program has limited materials--you should abandon an experimental test or run the experiment several times with different groups each time and then combine results to perform a single analysis.

What Program Should the Control Group Receive?

DECISION

If one group of students will get the program and a control group will not, the question arises about exactly what should happen to the control group. Should the control group receive no instruction in the subject matter to be taught by the program? For example, if program students leave the classroom to work on computer assisted instruction in fractions, should the control students receive instruction in fractions as well, or should they spend their time on something else altogether?

It is best to set up the experiment to match the way in which the program will be used in the future. If the program will be used as an adjunct to regular instruction, then set up the experiment so that the experimental group gets the program in addition to the regular program. If the program, on the other hand, is a replacement for regular instruction, then the control group will get only regular instruction and the experimental group will get only the program. If you are interested in assessing the effectiveness of two separate programs, either of which might replace the regular one, then give one to the experimental group and one to the control.

How To Design a Program Evaluation discusses what should happen to control groups on pages 29 to 32.

What Outcome Measure—Posttest—Is Reasonable for Detecting the Effect of One Cycle of the Experiment?

DECISION The posttest must meet the requirements of a good test. It should therefore be:

- Adequately long to have good reliability
- Representative of all the relevant objectives of the program, to demonstrate content validity
- Clearly understandable to the students

A good posttest is essential. Whether you plan to purchase it or construct it yourself, refer to How To Measure Achievement.

Do You Need Other Measures in Addition to the Outcome Measure?

DECISION Will the posttest provide a sufficient basis on which to judge the program? If the posttest contains many items which reflect specific details of the program—special vocabulary, for instance, or math problems that use a particular format—then a high posttest score may not represent much growth in general skills. In such a case, you might want to use an additional posttest for measuring achievement that contains more general items.

Since an immediate posttest will measure the initial impact of a program, you may wish to measure retention by administering another test some time later. You may, in addition, need to measure other program outcomes such as the attitudes of students, parents, or teachers.

See How To Measure Achievement and How To Measure Attitudes

What Procedures Will Be Used for Measuring Program Implementation?

DECISION As the program runs through a cycle, a record should be kept of which students actually participated in the program and which students—perhaps because of absences—did not. You must also keep careful track of what the experiences of program and control students looked like.

See How To Measure Program Implementation.

Which Groups of People Will Be Informed About the Results?

CHECK Check relevant audiences:

☐ Teachers of students involved
☐ The program's planners and curriculum designers
☐ Other teachers
☐ Principals
☐ District personnel
☐ Parents of students involved
☐ Other parents
☐ Board members
☐ Community groups
☐ State groups
☐ The media
☐ Teachers' organizations

Do meetings need to be held with any of these groups, either to give information or to hear their concerns, or for both reasons?

☐ Yes ☐ No

If yes, hold such meetings.

GO BACK You and the others involved have now finished deciding how to do the evaluation. Once these decisions are firm, go back to Step 1 and change the preconditions entries you made there if necessary.

Step 3

Record the Evaluation Plan

Instructions

Construct and complete a worksheet like the one below, summarizing the decisions made during Step 2. Contents of the worksheet can be used later as a first draft of parts of the evaluation report.

If two programs or components are being compared, and each is equally likely to be adopted, then you will have to carefully describe both.

PROGRAM DESCRIPTION WORKSHEET

This worksheet is written in the past tense so that when you have completed it you will have a first draft of two sections of your report: those that describe the program and the evaluation. For more specific help with deciding what to say, consult How To Present an Evaluation Report.

Background Information About the Program

A. Origin of the Program

B. Goals of the Program

C. Characteristics of the Program--materials, activities, and administrative arrangements

D. Students Involved in the Program

E. Faculty and Others Involved in the Program

Purpose of the Evaluation Study

A. Purposes of the Evaluation

4-5

B. Evaluation Design

A pretest-posttest true experiment was used to assess the impact of the program on student achievement. The target sample consisted of all who (fill in the selection criteria here)

Experimental and control groups were formed by random selection from pairs of students matched on the basis of the pretest.

C. Outcome Measures

D. Implementation Measures

Once you have completed the Worksheet, you have prepared descriptions of the program and of the evaluation. These descriptions will serve as your first draft of the evaluation report.

Step 4

Prepare or Select the Tests

Instructions

The Pretest

Use one of three kinds of pretests:

- A test to identify the sample of students eligible for the program--this is a selection test
- A test of ability given because you believe ability will affect results, and you therefore want the average abilities of the experimental and control groups to be roughly equal
- A pretest which is the same as the posttest, or its equivalent, so that you can be sure that the posttest shows a gain in knowledge that was not there before

In most cases, the pretest should be the posttest or the outcome measure itself. If this will be possible in your situation, then produce a thorough test which will be used as both pretest and posttest.

Preparing the Pretest Yourself

How To Measure Achievement, Chapter 3, lists resources, item banks, and guides to help you construct a test yourself. How To Measure Attitudes gives step-by-step directions for constructing attitude measures of all sorts.

Once the test has been written, try it out with a small sample of students to ensure that it is understandable and that it yields an appropriate pattern of scores for a pretest--not too many high scores so that there is room at the top for students to show growth. The tryout students should not be students who will be assigned to either the experimental or control groups. You will need at least five students for the tryout. They should be as similar as possible to the students who are to receive the program. You might need to borrow students from another class or school.

CHECK ✓ ☐

Check off these substeps in test development as you accomplish them:

☐ Test has been drafted or selected
☐ Test has been tried out with a small group of students
☐ Results of the tryout have been graphed and examined. Consult Worksheet 2A of How To Calculate Statistics for help with graphing scores.
☐ Test has been revised, if necessary
☐ Test has been reproduced in quantity ready for use

NOTE

If you intend to use the pretest you have purchased or written for selection of students, then you will, of course, have to administer the test before you decide which students are eligible. In this case, complete Step 6 before Step 5.

If the pretest will be administered to program and control groups after the groups have been formed, then go on next to Step 5.

Step 5

Prepare a List of Students

Instructions

List all students for whom one cycle of the program will be appropriate. In order to construct this list, you must have a set of criteria for selection. These should have been established in Step 1 and recorded on the worksheet in Step 3.

Write the names of the students who meet the selection criteria down the left hand side of the paper. Call this a sample list.

If you are using the selection test as a pretest as well, list students in order by score, from highest to lowest, and record each student's score next to his or her name.

Your sample list might look like this:

SAMPLE LIST

Adams, Jane
Bellows, John
Cartwright, Jack
Dayton, Maurice
Dearborn, Fred

Eaton, Susie
James, Alice
Markham, Mark
Payne, Tom
Pine, Judy

Taylor, Harvey
Vine, Grace
Washington, Roger
Williams, Greg

Step 6

Give the Pretest

Instructions

It is best to give the pretest at one sitting to all students concerned. Be sure no copies of the test are lost. All tests handed out must be returned at the end of the testing period. For obvious reasons, this is critical if the test will be used again as a posttest.

Tests are more likely to get lost when they use a separate answer sheet which is also collected separately. If your test uses a separate answer sheet, then have students place answer sheets inside the test booklet, and collect the two together.

4-8

Step 7

Form the Experimental and Control Groups

Instructions

a. Record pretest scores on the sample list if you have not already done so.

b. Graph the pretest scores

Refer to Worksheet 2A of How To Calculate Statistics for help with this step.

Are the scores appropriate for a pretest? That is, are scores relatively spread out with few students achieving the maximum? If yes, continue.

If the test was too easy, prepare and give another test with more difficult items. The program's instructional plans might need revision too if a test well-matched to the program's objectives was too easy for the target students.

c. Rank order the students according to pretest scores

If it is not already arranged according to student scores, rewrite the sample list starting with the student with the highest score and working down to the lowest.

d. Form "matched" pairs

Draw a line under the top two students, the next two, and so on.

Bellows	38
Eaton	36
Adams	35
Dayton	35
James	35
Payne	32
Dearborn	31
Vine	30

e. From each pair, randomly assign one student to the experimental group and the other student to the control group

To accomplish the random assignment, toss a coin. Call the experimental group or E-group "heads" and the control or C-group "tails." If a toss for the first person in the first pair gives you heads, assign this person to the E-group by putting an E by his name. His match, the other person in the pair, is then assigned to the C-group. If you get tails, the first person in the pair goes to the C-group and the other to the E-group.

Repeat the coin toss for each pair, assigning the first person according to the coin toss and his match to the other group. If there is an odd number of students, just randomly assign the odd student to one or the other group, but do not count him in the analysis later.

f. Prepare a Data Sheet

Have a list of the E-group and C-group students typed on a Data Sheet. This sheet should place the E-group at the left-hand side with a column for the posttest scores, then the C-group and the score column at the right. Always keep matched pairs on the same row. Columns 5, 6, and 7 will contain calculations to be performed later.

DATA SHEET

1 E-group	2 Post-test	3 C-group	4 Post-test	5 d	6 $(d-\bar{d})$	7 $(d-\bar{d})^2$

Step 8

Arrange For Your Implementation Measures

Instructions

Ensure that the program has been implemented as planned. This means ensuring that the students who are supposed to get the program--the E-group--do get it, and the others--the C-group--do not.

To accomplish this, try the following:

- Work closely with teachers to assure that the program groups receive the program at the appropriate times. Arrange a plan for carefully monitoring student absences from the program.
- Set up a record-keeping system to verify implementation of the program. For example, students could sign a log book as they arrive for the program, or perhaps they could turn in their work after each session. In addition, if possible, plan to have observers record whether the program in action looks the way it has been described.

GO BACK Refer back to the worksheet in Step 3 (Implementation Measures) to review your decisions on how to measure program implementation.

Check How To Measure Program Implementation for suggestions about collecting information to describe the program.

4-10

Step 9

Run the Program One Cycle

Instructions

Let the program run as naturally as possible, but check that accurate records are kept of the students' exposure to the program.

NOTE Be careful. If teachers or the evaluator pay extra attention to experimental group students, this alone could cause superior learning from them. So be as unobtrusive as possible.

Step 10

Posttest the E-group and C-group

Instructions

Give the posttest to the experimental and control groups at one sitting, if possible, so that testing conditions are the same for all students. If one sitting is not possible, test half the experimental group along with half the control group at one sitting and the others at a second sitting.

Of course, some of your outcome measures might not be tests as such. Interviews, observations, or whatever, should also be obtained from the experimental and control groups under conditions that are as similar as possible.

If necessary, schedule make-up tests for students absent from the posttest.

Step 11

Analyze the Results

Instructions

a. Score the Posttests

If the test you have constructed yourself contains closed response items—for example, multiple choice, true-false—then you can delegate someone to score the tests for you. How To Measure Achievement, pages 117 to 120, contains suggestions for scoring and recording results from your own tests.

b. Check the Data Set and Prune as Necessary

Use the Sample List to complete this procedure:

Check for absences from the program. If some students in either the experimental or control group missed a lot of school during the program's experimental cycle, they they should be dropped from the sample. You and your audience will have to agree about how many absences will require dropping the student from the analysis. One day's absence in a cycle of one week would probably be significant since it represents 20% of program time. A week's absence in a six month program, on the other hand, could probably be ignored.

If you decide that students in the experimental group should be dropped from the analysis if their absences exceeded, say, six days during the program, then control group students absent six or more days should also be dropped. This keeps the two groups comparable in composition. If the control group received a program representing a critical competitor to the program in question, then control group absences should be noted as well and the Sample List pruned accordingly.

From attendance records, determine the number of days each student was absent during the program cycle. Record this information in appropriately labeled columns added to the Sample List. Drop all students whose absences exceeded a tolerable amount for inclusion in the experiment. For every student dropped, the corresponding control group match will have to be dropped also. Drop as well any student for whom there is no posttest score. Drop his match also.

c. Summarize Attrition

Summarize results from pruning of the data in the table below. The number dropped from each group is called its "mortality" or "attrition."

TABLE OF ATTRITION DATA
Number of Students Remaining in the Study
After Attrition for Various Reasons

	Experimental Group	Control Group
Number assigned on basis of pretest		
Number dropped because of excessive absence from school during program		
Number dropped from E-group because of failure to receive program although in school		
Number dropped because of lack of posttest score		
Number dropped because match was dropped		
Number retained for analysis		

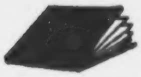
d. Record Posttest Scores on the Data Sheet for Students Who Have Remained in the Analysis

4-12

Instructions

e. Test To See if the Difference in Posttest Scores Is Significant

Were you to record just any two sets of posttest scores, it is likely that one of the groups would have higher scores than the other just by chance. What you now need to ask is whether the difference you will almost inevitably find between the E- and C-group posttest scores is so slight that it could have occurred by chance alone.



The logic underlying tests of statistical significance is described in How To Calculate Statistics. In fact, pages 71-76 of that book discuss the t-test for matched groups, to be used here, in detail.

To decide whether one or the other has scored significantly higher in this situation, you will use a correlated t-test--correlated because of the matched pairs used to form the two groups. Using your data, you will calculate a statistic, t. You will then compare this obtained value of t with values in a table. If your obtained value is bigger than the one in the table, the tabled t-value, then you can reject the idea that the results were just due to chance. You will have a statistically significant result. Below are the steps for this procedure.

Steps for Calculating and Testing t

Calculate t

This is the formula for t:

$$t = \frac{(\bar{d})\,(\sqrt{n})}{s_d}$$

In order to calculate it, you need to first compute the three quantities in the formula:

$\bar{d}$ = average difference score

$\sqrt{n}$ = the square root of the number of matched pairs

s_d = the standard deviation of the difference scores

Use the data sheet from Step 7 to help you calculate quantities for the t equation.

DATA SHEET

1 E-group	2 Post-test	3 C-group	4 Post-test	5 d	6 $(d-\bar{d})$	7 $(d-\bar{d})^2$

Page 126 shows a data sheet that has been computed.

To compute $\bar{d}$. First find the difference between the scores on the posttests for each pair of students. The difference, d, for a pair is the quantity:

[posttest score of the E-group student] - [posttest score of the matched C-group student]

Note that whenever a C-group student has scored higher than an E-group student, the difference is a negative number. Record these differences in Column 5 of the Data Sheet.

Then add up the entries in Column 5 and divide that sum by the number of pairs being used in the analysis, n. This gives you the average difference between the E-group and C-group. Call it $\bar{d}$, read "d bar."

$\boxed{\qquad} = \bar{d}$

To compute s_d. Fill in the quantities for Columns 6 and 7. For Column 6, subtract $\bar{d}$ from each value in Column 5, and record the result. For Column 7, square each number in Column 6 and divide their sum by n-1, the number that is one less than the number of pairs. Take the square root of your last answer and record this below as s_d.

$\boxed{\qquad} = s_d$

To compute $\sqrt{n}$. Take the square root of the number of matched pairs--not the number of students--which you are using in the analysis. This $\sqrt{n}$. Enter it here:

$\boxed{\qquad} = \sqrt{n}$

BEST COPY 4-13

Instructions

To compute t. Now enter these values in the formula for t below:

$$t = \frac{(\bar{d})\ (\sqrt{n})}{s_d} =$$

Multiply the top line. Then divide the result by s_d to get your t-value. Enter it here:

☐ = obtained t-value

Find the Tabled t-value

Using the table below, go down the left-hand column until you reach the number which is equal to the number of matched pairs you were analyzing. Be careful to use the number of pairs, not the number of students.

Table of t-Values for Correlated Means

Number of matched pairs	Tabled t-value for a 10% probability (one-tailed test)
6	1.48
7	1.44
8	1.41
9	1.40
10	1.38
11	1.37
12	1.36
13	1.36
14	1.35
15	1.34
16	1.34
17	1.33
18	1.33
19	1.32
20	1.32
21	1.32
22	1.32
23	1.32
24	1.32
25	1.31
26	1.31
.	.
40	1.30
.	.
120	1.29

The t-value in the left-hand column that corresponds to the number of matched pairs is your tabled t-value. Enter it here:

☐ = tabled t-value

Interpret the t-test

If the obtained t-value is greater than the tabled t-value, then you have shown that the program significantly improved the scores of students who got it. If your obtained t-value is less, then there is more than a 10% chance that the results were just due to chance. Such results are not usually considered statistically significant. The program has not been shown to make a statistically significant difference on this test.

The test of statistical significance which you have used here allows a 10% chance that you will claim a significant difference when the results were in fact only due to chance. If you want to make a firmer claim, use the Table of t-values in Appendix B. This table allows only a 5% chance of making such an error.

A good procedure in any case is to repeat the program another cycle and again perform this evaluation-by-experiment—only this time, use the 5% table to test the results. If your results are again significant, you will have very strong grounds for asserting that the program makes a statistically significant difference in results on the outcome measure.

Construct a Graph of Scores

If results were statistically significant, display them graphically. Figures A and B present two appropriate ways to do this. Figure A requires fewer calculations.

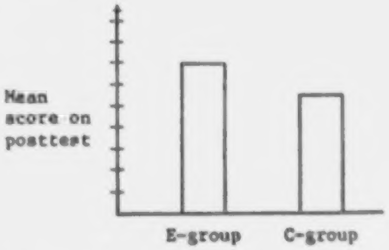


Figure A. Posttest means of groups formed from matched pairs

Instructions

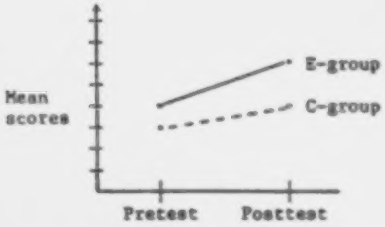


Figure B. Pretest and posttest mean scores of experimental and control groups

You may wish to take a closer look at the results than just examining averages or single tests of significance. Taking a close look will further help you interpret results. In particular, if the results were not statistically significant, you may want to look for general trends.

One good way to take a closer look at results is to compute the gain score--posttest minus pretest--for each student. Using gain scores, you can plot two bar graphs, one showing gain scores in the experimental group and the other showing gain scores in the control group. If some students' scores were quite extreme, look into these cases. Perhaps there was some special condition, such as illness or coaching, which explains extreme scores. If so, these students' scores should be dropped and the t-test for differences in posttest scores computed again.

BEST COPY

4-15

Step 12

If Your Task Is Formative, Meet With the Staff To Discuss Results

Instructions

The agenda for this meeting should have an outline something like the following:

Introduction

Review the contents of the worksheet in Step 3, pages 111 and 112.

Presentation of Results

Display and discuss the attrition table which describes student absences from the experimental and control groups. Display Figures A and B and discuss them. Report the results of the test of significance.

Discussion of the Results

If the difference was significant as hypothesized—the E-group did better than the C-group—you will need to answer these questions:

- Was the result educationally significant? That is, was the difference between the E-group and the C-group large enough to be of educational value?
- Were the results heavily influenced by a few dramatic gains or losses?
- Were the gains worth the effort involved in implementing the program?

If the results were non-significant, you will need to consider:

- Do you think this was due to too short a time span to give the program a fair chance to show its effects, or was the program a poor one?
- Were there special problems which could be remedied?
- Was the result nearly significant?
- Should the program be tried again, perhaps with improvements?

Recommendations

On the basis of the results, what recommendations can be made? Should the program be expanded? Should another evaluation be conducted to get firmer results—perhaps using more students? Can the program be improved? Could the evaluation be improved? Collect and discuss recommendations.

4-14

BEST COPY

Step 13

Write a Report if Necessary

Instructions

Use as resources the book How To Present an Evaluation Report and the worksheet in Step 3 of this guide. The worksheet, you will remember, contains an early draft of the sections of the report that describe the program and the evaluation.

You have reached the end of the Step-by-Step Guide for Conducting a Small Experiment. The guide, however, ~~has two appendices:~~

- ~~Appendix A~~ contains in Appendix A an example of an evaluation report prepared using this guide.
- ~~Appendix B contains the table of values for performing a t test of statistical significance at the 5% level.~~

4-17

Appendix A

Example of an Evaluation Report

This example--which is fictitious and should not be interpreted as evidence for or against any particular counseling method--illustrates how an experiment can form the nucleus of an evaluation. Notice that information from the experiment does not form the sole content of the report. The evaluator has to consider many contextual, program-specific pieces of information, such as the exact nature of the program, the possible bias that might be introduced into the data by the information available to the respondents, etc. There is no substitute for thoughtfulness and common sense in interpreting an evaluation.

EVALUATION REPORT

Program	The Preventive Counseling Program
Program location	Naughton High School
Evaluator	J. P. Simon, Principal Naughton High School
Report submitted to	J. Ross, Director of Evaluation Mimieux School District
Period covered by report	January 6, 19xx-February 16, 19xx
Date report submitted	March 31, 19xx

Section I. Summary

A new counseling technique based on "reality therapy" and the motto that "prevention is better than cure" was developed by the Mimieux School District and consultants.

Naughton High School evaluated this Preventive Counseling Program by making it available to one group of students, but not to a matched control group.

Results of teacher ratings subsequent to the Preventive Counseling Program and a count of the number of referrals to the office, both pointed to the success of the PC program at least on this short-term basis.

This evaluation report details these findings and presents a series of recommendations for further evaluation of this promising program.

Section II. Background Information Concerning The Preventive Counseling Program

A. Origin of the Program

Several counselors had received special training, at district expense, in a style of counseling related to "reality therapy." This counseling was designed to be used with students whom teachers felt were "heading for trouble" in school or not adjusting well to school life. By an intensive course of counseling, it was hoped to prevent future problems, hence the title the Preventive Counseling Program. The district office asked Naughton High School to assess the effectiveness of this kind of counseling. A counselor trained in the technique was made available to the school on a trial basis for four hours a day over a two week period.

B. Goal of the Program

The goal of the Preventive Counseling Program (PC) was to promote successful adjustment to school among students whom teachers referred to the office.

C. Characteristics of the Program

In the PC program, a student who is referred by a teacher receives an initial 20 minutes of counseling. Follow-up counseling sessions are given to the student each day for the next two weeks.

This program differs from methods used previously to handle referrals to the office. Previously, teachers were not encouraged to refer students to the office. When a student was referred for some

4.18

particular reason, he generally received one counseling session and perhaps no follow-up at all, unless the teacher referred the student again. This kind of counseling was the responsibility of the usual counseling staff or, in exceptional cases, the vice-principal.

The PC program:

1. Uses counselors who are specially trained in "reality therapy" counseling
2. Requests referrals before an incident necessitates referral
3. Gives the student two weeks of counseling

D. Students Involved in the Program

The counseling is appropriate for students of all grade levels. Any student referred by a teacher is eligible for counseling. During the trial period for this evaluation, however, only some referred students could receive the PC program.

E. Faculty and Others Involved in the Program

As far as possible, the counselor and teachers communicated directly regarding students in need of counseling. A clerk handled scheduling of counseling sessions, managing this in addition to his other duties.

Section III. Description of the Evaluation Study

A. Purposes of the Evaluation

The District Office wanted Naughton High School to evaluate the effectiveness of the new style of counseling. The study in this school was to be one of several studies conducted to assist the District in deciding whether or not to have other counselors receive reality therapy training and conduct preventive counseling.

Several School Board members had emphasized that they were interested in seeing firm evidence, not opinions.

B. Evaluation Design

In view of the costly decisions to be made and the desire of the Board members for "hard data," the evaluation was designed to measure the results of the PC program as objectively and accurately as possible. To accomplish this, it was deemed necessary to use a true control group. Teachers were asked to name students in their classes who were in need of counseling. For each student named, the teacher provided a rating of the student's adjustment to school on a 5-point scale from "extremely poor" to "needs a little improvement." This was called the adjustment rating.

Students referred by three or more teachers formed the sample used in the evaluation. An average adjustment rating was calculated for each of the sample students by adding together all ratings for a student and dividing by the number of ratings for that student. These students were then grouped by grade and sex. Matched pairs were formed by matching students (within a group) with close to the same average ratings.

From these matched pairs, students were randomly assigned to receive the new counseling (the Experimental or E-group) or to be the Control group or C-group. Should students from the control group be referred for counseling because of some incident, for example, then the regular counselors were requested to counsel as they had in the past. The E-group students received the two weeks of counseling which is characteristic of the PC program.

At the end of the two-week cycle, all referrals to the office were again dealt with by regular counselors or the vice-principal. Over the next four weeks, records of referrals to the office were kept. If the number of referrals to the office was significantly fewer for the students who had received the PC program (i.e., the E-group students), then the program would be inferred to have been successful.

This measure is reasonably objective and the random assignment of students from matched pairs ensured the initial similarity of the two groups, thus making it possible to conclude that any difference in subsequent rates was due to the PC program.

C. Outcome Measures

As mentioned above, the effect of the program was measured by counting, from office records, how many times each control group student and how many times each experimental group student was referred to the office in the four weeks after the intervention program ended.

An unavoidable problem was that teachers were sometimes aware of which students had been receiving the regular counseling, since students were called to the office regularly for two weeks from their classes. Teachers might have been influenced by this fact. In order to reduce the possible impact of this situation on teacher referral behavior, the fact that the evaluation was being conducted was not made known until after the data collection period was over (four weeks after the Preventive Counseling program ended).

A second measure of outcomes was also collected: teachers were asked at the end of the data collection period to re-rate all students previously identified as needing counseling, giving a "student adjustment rating" on the same 5-point scale which had been used in the beginning of the program.

D. Implementation Measures

The counselor's records provided the documentation for the program. Essentially, these records were used to verify that only E-group students had received the Preventive Counseling program and to record any absences which might require that the student not be counted in the evaluation results.

Section IV. Results

A. Results of Implementation Measures

Eighteen pairs of students were formed from teachers' referrals. The 18 students in the E-group had a perfect attendance record during the Preventive Counseling program and did not miss any counseling sessions. However, two students in the control group were absent for a week. These students and their matched pairs were not counted in the analysis thus leaving a total of 16 matched pairs.

B. Results of Outcome Measures

Table 1 shows the number of referrals to the office from the experimental and control groups during each of the four weeks following the end of the PC program.

TABLE 1
Number of Referrals to the Office

	# of referrals to office				
	Week 1	Week 2	Week 3	Week 4	Total
E-group (had received PC)	1	1	1	2	5
C-group (had not received PC)	3	2	3	2	10

There were twice as many referrals (10 as opposed to 5) in the control group as in the experimental group. Closer analysis revealed that four of the referrals in the E-group were produced by one student who was referred to the office each week. Checking the number of students referred at least once (as opposed to the total number of referrals), it was found that there were two for the experimental and six for the control group.

The second set of averaged school adjustment ratings collected from teachers is recorded in Figure 1, and the calculations for a test of the significance of the results are presented in the same figure. The t-test for correlated means was used to examine the hypothesis that the E-group's average adjustment ratings would be higher, after the program, than those of the C-group. The hypothesis could be accepted with only a 10% chance that the obtained difference was simply the result of chance sampling fluctuations. The obtained t-value was 2.06, and the tabled t-value (.10 level) was 1.34.

DATA SHEET

E-group		C-group				
Student	Final average adjustment rating	Student	Final average adjustment rating	d	$(d-\bar{d})$	$(d-\bar{d})^2$
AK	3	WK	1	2	1.38	1.90
GF	2	LJ	2	0	- .62	0.38
ST	4	CF	1	3	2.38	5.66
CT	4	LM	3	1	0.38	0.14
JB	3	MH	3	0	-0.62	0.38
SK	3	FH	4	-1	-1.62	2.62
UL	5	DH	5	0	-0.62	0.38
MQ	5	RR	4	1	0.38	0.14
JJ	3	XT	1	2	1.38	1.90
WV	2	KN	2	0	- .62	0.38
AC	4	JR	3	1	0.38	0.14
CK	3	OF	4	-1	-1.62	2.62
CR	2	PD	1	1	0.38	0.14
RA	5	NW	5	0	-0.62	0.38
PG	3	JM	4	-1	-1.62	2.62
FW	4	RL	2	2	1.38	1.90
n = 16				10		21.68

$$\bar{d} = \frac{13-3}{16} = \frac{10}{16} \qquad s_d = \sqrt{\frac{21.68}{15}} = \sqrt{1.44}$$

$$\boxed{\sqrt{n} = 4} \qquad \boxed{\bar{d} = 0.62} \qquad \boxed{s_d = 1.20}$$

$$t = \frac{(\bar{d})\ (\sqrt{n})}{s_d}$$

$$t = \frac{(0.62)\ (4)}{1.20} = \frac{2.48}{1.20} = \boxed{2.06}$$

Figure 1

4-20

C. Informal Results

Several teachers commented informally about the counseling that their problem students were receiving. One said the counseling seemed to be less "touchy feely" and more "getting down to specifics," and she noted an increase in task orientation in a counselee in her room beginning at about the second week of special counseling. She felt, however, that the counseling should have continued longer. Other teachers did not seem to have ascertained the style of counseling being used, but commented that counseling seemed to be having less transitory effect than usual.

A parent of one of the counselees in the PC program called the principal to praise the consistent help his child was getting from the special counselor. "I think this might turn him around," the parent said.

Negative comments came from one teacher who complained that one of her students always seemed to miss some important activity by being summoned to the counseling sessions. Another teacher, however, commented that it was a relief to have the counselee gone for a little while each day.

Section V. Discussion of Results

The use of a true experimental design enables the results reported above to be interpreted with some confidence. Initially, the E-group and C-group were composed of very similar students because of the procedure of matching and random assignment. The E-group received preventive counseling whereas the C-group did not. In the four weeks following the program, all students were in their regular programs and during this time, students from the C-group received twice as many referrals to the office as students from the E-group.

In interpreting this measure, it should be remembered that referral to the office is a quite objective behavioral measure of the effect of the program. It appears that the Preventive Counseling program substantially reduced the number of referrals to the office over this four week period. Whether this difference will continue is not known at this time.

The average post-counseling ratings which teachers assigned to students in the E-group and in the C-group showed a significant difference in favor of the E-group. A problem in interpreting this result is that the teachers were aware of which students had been in the counseling program and this might have affected their ratings. However, 52 teachers were involved in these ratings, some rating only one student and others rating more. That the result was in the same direction as the behavioral measure lends both measures additional credibility.

Section VI. Cost-Benefit Considerations

The program appears to have an initially beneficial effect. However, it also is a fairly expensive program. There are two main expenses involved: the cost of training counselors in reality therapy and the cost of providing the counseling time in the school. There was no way in this evaluation of determining if the training had an important influence on the program's effectiveness. It could have been that other program characteristics--its preventive approach or the continuous daily counseling--were the influential characteristics. Training in reality therapy could possibly be dispensed with thus saving some of the expense. However, since training can presumably have lasting effects on a counselor, its cost over the long-run is not great and comes nowhere near approaching the cost of the provision of counseling time each day.

It is understood that a cost-benefit analysis will be conducted by the District office using results from several schools. One question needing consideration is whether the Preventive Counseling program will in fact save personnel time in the long run by catching minor problems before they develop into major problems. To answer such a question requires the collection of data over a longer time period than the few weeks employed in this evaluation. If the program helps students to overcome classroom problems, then its benefits--although perhaps immeasurable--might be great.

Section VII. Conclusions and Recommendations

A. Conclusions

In this small scale experiment, the Preventive Counseling program appeared to be superior to normal practice. It produced better adjustment to school, as rated by teachers, and resulted in fewer teacher referrals to the office in the four weeks following the end of the two week PC program. It was not possible to determine, from this small study, the extent to which each of the program's main characteristics was important to the success of the overall program.

B. Recommendations Regarding the Program

1. The Preventive Counseling program is promising and should be continued for further evaluation.
2. Preventive Counseling without the reality therapy training might be instituted on a trial basis.

C. Recommendations Regarding Subsequent Evaluation of the Program

1. The kind of evaluation reported here, an evaluation based on a true experiment and fairly objective measures, should be repeated several

BEST COPY 4-21

times to check the reliability of the effects of counseling as so measured.

2. In several evaluations of the Preventive Counseling program, the outcome data should be collected over a period of several months to assess long-term effects.

3. The School Board and the schools should be provided with a cost analysis of the counseling program which includes a clear indication of (a) the alternative uses to which the money might be put were it not spent on the PC program, and (b) the cost of other means of assisting students referred by teachers.

4. An evaluation should be designed to measure the relative effectiveness of the following four programs:
   - The Preventive Counseling program
   - The Preventive Counseling program run without reality therapy training
   - Reality therapy provided to regular counselors
   - The usual means of handling referrals

4.22

HOW TO DEFINE YOUR ROLE AS AN EVALUATOR

(DRAFT)

Revision Author: Brian Stecher

December 6, 1985

How To Focus An Evaluation

Brian Stecher

Outline

Introduction

A. Purposes of the book.

B. What it will and will not tell you.

C. Chapter by chapter overview.

Chapter One: Presenting a model for focusing an evaluation.

A. Preliminary comments/caveats.

   1. Limitations of any model of complex human interactions.

   2. Value as a tool for learning and instruction.

B. What are the elements of the focusing process?

   1. Acknowledging existing beliefs and expectations.

      a. Evaluator has beliefs about the meaning of evaluation, embodied in a particular approach.

      b. Client has expectations for the evaluation, based upon needs and wants.

   2. Gathering information.

      a. Evaluator seeks information about many topics.

         (1) Client's needs and expectations.

         (2) Program goals and activities.

         (3) Other concerned individuals or groups.

         (4) Constraints and limitations, etc. '

  - b. Client seeks information about many topics.
    - (1) Evaluator's capabilities.
    - (2) Value and limitations of evaluation.
    - (3) Evaluation procedures, etc.
- 3. Narrowing the focus and formulating a tentative strategy.
  - a. Establishing priorities.
  - b. Formulating preliminary plans.
  - c. Melding the evaluator's approach and the client's expectations.
- 4. Negotiating an evaluation plan.
  - a. Specifying evaluation questions.
  - b. Clarifying procedures.

Chapter Two: Thinking about client concerns and evaluator approaches.

- A. Client needs and expectations.
  - 1. Why consult an evaluator?
    - a. Legal mandates.
    - b. Stated program goals and objectives.
    - c. Specific questions.
    - d. General concerns or problems.
  - 2. Client conception of evaluation.
- B. There are different approaches to evaluation.
  - 1. What we mean by an "evaluation approach."
  - 2. Derivation of these points of view.
- C. The research approach.
  - 1. Conception of the meaning and purposes of evaluation.
  - 2. Methods for accomplishing these purposes.

D. The goal-oriented approach.
    1. Conception of the meaning and purposes of evaluation.
    2. Methods for accomplishing these purposes.

E. The decision-focused approach.
    1. Conception of the meaning and purposes of evaluation.
    2. Methods for accomplishing these purposes.

F. The user-oriented approach.
    1. Conception of the meaning and purposes of evaluation.
    2. Methods for accomplishing these purposes.

G. The responsive approach.
    1. Conception of the meaning and purposes of evaluation.
    2. Methods for accomplishing these purposes.

H. Comparison of approaches.
    1. Similarities: information, validity, usefulness, etc.
    2. Differences: research paradigm, degree of subjectivity, role of the evaluator, etc.

Chapter Three: How to gather information.

A. Introduction: This is a simplified discussion of a complex, interactive procedure.
    1. It is a dynamic process that differs in each case.
    2. As the expert the evaluator is likely to have strong influence.
    3. There are fundamental concerns common to all evaluators can be captured in four or five basic questions.

B. "What is the program all about?"
    1. Obtaining information about the program.
    2. How different evaluators might ask this question.
        a. What variables do you want to study?
        b. What are your goals and objectives?
        c. What decisions are going to be made?
        d. Who is likely to use the information?
        e. Who is affected by the program?
    3. How different clients might respond.

C. "What do you want to know about the program?"
    1. Illustrations of various points of view.
    2. Questions each evaluator will want to have answered.

D. "Who else is concerned and may need to be involved?"
    1. Extending the information base.
    2. How different evaluators would address this issue?

E. "Why do you want this information?"
    1. Clients view of purposes for the evaluation.
    2. Impact on different evaluators.

F. "What constraints or limitations are there?"
    1. What practical limits exist: money, time, access, etc.?
    2. What contextual constraints exist: attitudes, politics, beliefs, etc.?
    3. How would different evaluators address these issues?

Chapter Four: How to narrow the focus and develop tentative plans.

1. Developing and revising plans as information is gathered.
2. Establishing priorities.

3. Adapting strategies to fit particular situations.
4. Balancing evaluator's point of view and client's wishes.

Chapter Five: How to negotiate an evaluation plan.

1. Desired outcomes.
    a. Specific objectives and evaluation questions.
    b. Methods and procedures.
2. Options for the evaluator.
    a. Reach a collaborative agreement.
    b. Decline to conduct the evaluation.

Closing Comments

1. Summary
2. Return to the Kit.

HOW TO DESIGN A PROGRAM EVALUATION

(DRAFT)

Revision Author: Joan Herman

December 6, 1985

Chapter 1

An Introduction to Evaluation Design

A design[1] is a plan which dictates when and from whom information is to be collected during the course of an evaluation. The first and obvious reason for using a design is to ensure a well organized evaluation study: all the right people will take part in the evaluation at the right times. A design, however, accomplishes for the evaluator something more useful than just keeping data collection on schedule. A design is most basically a way of gathering data so that the results will provide sound, credible answers to the questions the evaluation is to address.

The term design traditionally has been used in the context of quantitative evaluation studies where judgments of program worth or of relative effectiveness are a primary consideration. This book is addressed to designs for these types of studies. It is important to note, however, that there are occasions when a quantitative study does not represent the best approach to answering important evaluation questions, where qualitative approaches may be more appropriate. Design is equally important in assuring the quality of information derived from qualitative studies. The reader is referred to How to Conduct Qualitative Studies for a discussion of important design issues in these latter types of studies.

What is the purpose of design in quantitative studies? A design is a plan for gathering comparative information so that results from the program being evaluated can be placed within a context for judgment of their size and worth. Designs reinforce conclusions the evaluator can draw about the impact of a program by helping the evaluator to predict how things might have been had the program not occurred or if some other program had occurred instead.

Chapter 1

An Introduction to Evaluation Design

[illegible]

A *design*[1] is a *plan* which dictates *when* and *from whom* measurements will be gathered during the course of an evaluation. The first and obvious reason for using a design is to ensure a well organized evaluation study: all the right people will take part in the evaluation at the right times. A design, however, accomplishes for the evaluator something more useful than just keeping data collection on schedule. A design is most basically a way of gathering *comparative information* so that results from the program being evaluated can be placed within a context for judgment of their size and worth. Designs reinforce conclusions the evaluator can draw about the impact of a program by helping the evaluator to predict *how things might have been had the program not occurred* or if some other program had occurred instead. The comparative data collected could include how the school environment might have looked, how people might have felt, and how participants might have performed had they not encountered the particular program under scrutiny. Usually a design accomplishes this by prescribing that measurement instruments—tests, questionnaires, observations—be administered to comparison groups *not receiving the program.* These results are then *compared* with those produced by program participants. At other times, predictions about what would have happened in the program's absence can be produced without a comparison group through application of statistical techniques.

---

[1] Some writers have used the word *model* instead of design, probably because the choice of such a measurement plan usually affects the evaluator's whole point of view about the seriousness of the enterprise and about how information will be gathered, analyzed, and presented. This book prefers *design,* the less ponderous term, and it will be used throughout.

The objective of this book is to acquaint you with the ways in which evaluation results can be made more credible through careful choice of a design prescribing when and from whom you will gather data. The book helps you choose a design, put it into operation, and analyze and report the data you have gathered. The book's intended message is that attention to design is important.

Even if choice or practicality dictate that you ignore the issue of design, it is important that you understand the data interpretation options which you have chosen to pass by. In the majority of evaluation situations, *some* comparative information is better than none. Your choice of a design will perhaps determine whether the information you produce is believed and used by your evaluation audience or shrugged off because its many alternative interpretations render it unworthy of serious attention.

The book's contents are based on the experience of evaluators at the Center for the Study of Evaluation, University of California, Los Angeles, on advice from experts in the field of educational research, and on the comments of people in school settings who used a field test edition. The book focuses on those evaluation designs which seem most practical for use in program evaluation. Please be aware that these are *not the only* designs available for adoption as bases for useful research. They do seem to be, however, the most straightforward and intuitively understandable. This makes them likely to be accepted by the lay audiences who will receive and must interpret your evaluation findings. Please bear in mind, in addition, that many of the recommended procedures in this book prescribe the design of a program evaluation *under the most advantageous circumstances.* Few evaluation situations exactly match those envisioned here or described in the book's myriad examples. Therefore, *you should not expect to duplicate exactly suggestions in the book.* Evaluation is a relatively new field, and correct procedures, even where choice of a design is concerned, are not firmly established. In fact, while considerable attention has been given to the quality of measurement instruments for assessing cognitive and affective effects of programs, relatively little attention has been paid to the provision of useful designs. Your task as an evaluator is to find the design that provides the most credible information in the situation you have at hand and then to try to follow directions as faithfully as possible for its implementation. If you feel you'll have to deviate from the procedures outlined here, then do. If you think the deviation will affect interpretation of your results, then include the appropriate qualifications in your report.

If political pressures or the heat of controversy make it important that you produce credible information about program effects, few things will support you better than a well chosen evaluation design. Often evaluators discouraged by political or practical constraints have chosen to ignore design, perhaps cynically deciding that a good design represents information overkill in a situation where little attention will be paid to the data anyway. The experience of evaluators who have chosen to use good design has been to the contrary. The quality of information provided through use of design has often *forced* attention to program results. Without design, the information you present will in most cases be haunted by the possibility of reinterpretation. *Information from a well designed study is hard to refute*; and in situations where they might have been ignored or shrugged off because of many or ambiguous interpretations, conclusions from a good design cannot be easily ignored.

The *Program Evaluation Kit*, of which this book is one component, is intended for use primarily by people who have been assigned the role of *program evaluator*. The job of program evaluator takes on one of two characters, and at times both, depending upon the tasks that have been assigned:

1. You may have responsibility for producing a *summary statement* about the effectiveness of the program. In this case, you probably will report to a funding agency, governmental office, or some other representative of the program's constituency. You may be expected to describe the program, to produce a statement concerning the program's achievement of announced goals, to note any unanticipated outcomes, and possibly to make comparisons with alternative programs. If these are the features of your job, you are a *summative evaluator*.
2. Your evaluation task may characterize you as a helper and advisor to the program planners and developers or even as a planner yourself. You may then be called on to look out for potential problems, identify areas where the program needs improvement, describe and monitor program activities, and periodically test for progress in achievement or attitude change. In this situation, you are a "jack of all trades," a person whose overall task is not well defined. You may or may not be required to produce a report at the end of your activities. If this more loosely defined job role seems closer to yours, then you are a *formative evaluator*.

The information about design contained in this book will be useful for both the formative and summative evaluator, although the perspective of each will vary.

Designs in Summative Evaluation

Typically, design has been associated with summative evaluation. After all, the summative evaluator is supposed to produce a public statement summarizing the program's accomplishments. Since this report could affect important decisions about the program's future, the summative evaluator needs to be able to back up his findings. He therefore has to anticipate the arguments of skeptics or even the outright attacks of opponents to the conclusions he presents. While good design won't immunize him against attack, it will strengthen his defense. Historically, designs were developed as methods for conducting scientific experiments, methods through which one can logically rule out the effect on outcomes of anything other than the *treatment* provided. In the case of educational evaluation, this treatment is an educational program. Since designs serve the interest of producing defensible results, and since such production is primarily the interest of the summative evaluator, you will find throughout the book a strong summative flavor in both the procedures outlined and the examples described.

To readers who are working right now as evaluators, the suggestion that *design* is of critical importance for summative evaluation may seem a little off-base. "No one uses experimental designs," you might say. "No one uses control groups." And you would be nearly correct, unfortunately—at least with regard to large Federal and State funded programs. Not long ago a study of a nationwide sample of ESEA Title VII (Bilingual Education) evaluations revealed that *no one* attempted to use a *true*, randomized control group, and only 36% tried to locate a non-randomized control group for comparison with any aspect of the programs evaluated.[2] In another study, a search of 2,000 projects that had received recognition as successful located not one with an evaluation that provided acceptable evidence regarding project success or failure.[3]

The reasons for this state of affairs are no doubt legion, but four come up frequently:

1. *Funders seem to view programs as one-shot enterprises.* Once a program has been implemented and has run its course, it becomes a *fait accompli.* It's over. Summative reports, then, describe something that has already happened. They are seldom seen as a chance to describe programs and their effects in the interest of future planning. In order to testify that a program took place *at all,* a summative report need not use a design. Designs become valuable only when someone hopes to use information about program processes and effects as a basis for future decisions such as whether to pay for similar programs or to expand the current one. Designs are *essential* when someone has in mind the development of theories about what instructional, management, or administrative strategies work best.

---

[2] Alkin, M. C., Kosecoff, J., Fitz-Gibbon, C., & Seligman, R. *Evaluation and decision making: The Title VII experience.* CSE Monograph Series in Evaluation, No. 4. Los Angeles: Center for the Study of Evaluation, 1974.

[3] Foat, C. M. *Selecting exemplary compensatory education projects for dissemination via project information packages.* Los Altos, CA: RMC Research Corporation, May, 1974 (Technical Report No. UR-242).

2. *Evaluators are called in too late.* This problem is actually a common symptom of the first. Evaluation often occurs as an afterthought. Lack of careful planning in the establishment of the program removes the possibility of a carefully planned evaluation. The evaluator finds that he has no control over the assignment of students or the sites chosen for implementation of the program. The evaluator has to "evaluate" an already on-going program. While this situation does not eliminate the possibility of obtaining good comparative information, it usually makes use of the best designs impossible.

Insert

3. Because of ethical and/or political concerns, it is often difficult to accomplish the most rigorous designs. Social programs often are aimed at individuals or groups in great need, and withholding potential program benefits from some for the sake of a comparative research design can be hard to justify. In addition, it is frequently the case that politics rather than social science methodology determines where or for whom special programs will be implemented, precluding opportunities for randomized designs.

4. *Social science research in general is still in its youth.* Lack of research design in evaluation stems partly from its relative novelty as a method for gathering social science information at all. Sir Ronald Fisher's work in statistics and design, an essential methodological step forward for the social sciences, was completed in the 1930's! Not very long ago.

5. *Educational researchers and evaluators themselves cannot agree about the appropriateness of research designs for evaluation.* While most writers in the field of evaluation concur that at least a part of the evaluator's role is to collect information about a program, the nature of the rules governing data collection are still debated. Opponents of the use of design usually list as major drawbacks the political and practical constraints discussed here already, and the technical difficulties involved with using the findings from one multifaceted program to predict the outcomes of others.

Defenders of design, the authors of this book among them, acknowledge these [illegible]. They continue to urge the use of design in field settings because designs yield the comparative information necessary for establishing a perspective from which to judge program accomplishments. In fields of endeavor such as education, where clear absolute standards of performance have not been set, comparison is a way to subject programs to scrutiny in order eventually to determine their value. Nonetheless, some of these impediments to good design are more intractable than others. Suggestions are offered at the end of this chapter for optimizing those situations where there are significant intractable constraints.

problems

Insert:

Summative Evaluation and Educational Research

Summative evaluations should whenever possible employ experimental designs when examining programs that are to be judged by their results. The very best summative evaluation has all the characteristics of the best research study. It uses highly valid and reliable instruments, and it faithfully applies a powerful evaluation design. Evaluations of this caliber could be published and disseminated to both the lay and research community. Few evaluations of course will live up to such rigid standards or need to. *The critical characteristic of any one evaluation study is that it provide the best possible information that could have been collected under the circumstances, and that this information meet the credibility requirements of its*

BEST COPY

evaluation audience. The best interpretation of your task as summative evaluator is that you must collect the most believable information you can, anticipating at all times how a skeptic would view your report. Keeping this skeptic in mind, set about designing the evaluation which has potential for answering the largest number of criticisms.

The aim of the *researcher* is to provide findings about a program which can be generalized to other contexts beyond it. Criteria for what constitutes *generalizable* information have been agreed upon by the social science community; they are the topic of educational research texts. Though it is important as a service to education that the *evaluator* provide such information if the situation allows good design and high quality instrumentation, the evaluator can usually limit his projection of the quality of data he must collect to what he perceives will be acceptable to his unique audience. It is not beyond the scope of the evaluator's job, however, to *educate his audience* about what constitutes good and poor evidence of program success and to *admonish them* about the foolishness of basing important decisions on a single study—or even a few. It is equally within the summative evaluator's task to *advocate*, based on his information, *changes* in a program or in funding policy or to *express opinions* about the program's quality. The evaluator who takes a stand, however, must realize that he will need to defend his conclusions, and this again means good data and a well designed study.

Designs in Formative Evaluation

All this discussion about design in summative evaluation should not persuade the evaluator that design is irrelevant in the formative case. The use of design during a program's formative period gives the evaluator, and through her the program staff, a chance to take a good hard look at the effectiveness of the program or of selected subcomponents. This enables the formative evaluator to fulfill one of her major functions—to persuade the staff to constantly scrutinize and rethink assumptions and activities that underly the program. Careful attention to design can also help the formative evaluator to conduct small-scale pilot studies and experiments with newly-developed program components. These will inform decisions among alternative courses of action and settle controversies about more or less effective ways to install the program.

The message to the formative evaluator is this: Including a source of comparative information—a control group or data from time series measures—in any information-gathering effort makes that information more interpretable. Too often formative measurement happens in a vacuum; no one can judge whether students are making fast enough progress, for instance, because no one can answer the question "Compared to what?"

Example 1. Franklin Elementary School has designed a pull-out program in reading for slow readers and wishes to assess the quality of the progress of the students during the program's first year of operation. The hope is that students in the pull-out program will make faster progress because of increased attention. The problem is knowing what pace to expect from slow students. The vice-principal, serving as program evaluator, has located a school in the same district which uses the same programmed readers which form the backbone of the pull-out program—the O'Leary Series. The evaluator has persuaded the principal of the other school to allow her periodically to test their slow readers for comparison with Franklin's. The evaluator has constructed a test using sample sentences from the O'Leary series which will be administered for oral reading by both the pull-out program students and the students in the other school. Since it is the first year of the pull-out program, information gained from comparing the two schools will be used formatively. If Franklin's readers are not progressing faster than the controls, then this might signal a need for modification in the pull-out program. The design here is Design 5, the Time Series with Non-Equivalent Control Group, described in Chapter 5.

Example 2. Osirus High School designed a six-week career awareness module for tenth grade based on field trips in which all students spend one afternoon a week at the work places of professionals pursuing careers in which the students are interested. The students conduct interviews and write short biographies describing each professional's route to success. Due partly to the extreme cost of such a large-scale field program, the school's director of vocational education decided to do some formative evaluation, assigning students randomly to the first six-week program tryout. This provided a Design 2 evaluation (Chapter 4), since no pretest was given. At the end of the six-week module, an achievement test revealed that students had acquired large amounts of information about the careers of their choice, and were able to write essays which the career education staff judged to be realistic appraisals of the economic and social accompaniments to these careers. Students also seemed to have acquired a good sense of the steps necessary to attain an education toward the career of interest. A look at the control group, however, showed that students who had *not* taken part in the career education program had acquired the same information and the same set of realistic expectations simply through talking to students who were taking part in the field test. It seemed that it might not be necessary for every student to go into the field every week—at least this didn't seem critical for making cognitive gains.

For formative evaluation, it is a good idea at the outset to locate or assemble a control group, as described in Designs 1, 2, and 3, or to collect time series measures before the program begins (Designs 4 and 5). Laying down even the rudiments of design will give you a chance to make comparisons in order to interpret your findings or to justify your formative recommendations if you should need to.

nature Because of the ~~informality~~ of your job, you can try using designs for formative evaluation in several ways, according to your own discretion:

1. *You might set up as "controls" various alternative versions of the program you are helping to form.* You may be able to identify alternative versions that the program can take, possibly one or more less costly or time-consuming than the others. You could set up two or more versions in different schools or classrooms, some receiving the more expensive or more lengthy alternative. These alternatives could vary in the amount they differ from the basic program, as well as in their duration. They could last a short time, say, until someone has determined their relative quality; or they could span the duration of the whole program, providing you at the end with an assessment of their—and its—overall effectiveness. If whole schools or classrooms received the alternative version, your evaluation would comprise a Design 3 study (Chapter 4), an evaluation with a non-equivalent control group. If you have programs going on in several different classrooms to which you can randomly assign students, you can implement a true control group design. Often the "control group" tends to be thought of as a set of losers, people who unluckily miss out on all the good benefits of the program. In a design which sets two competing versions of the program operating at the same time and where each of them is equally viable and potentially effective, exactly which group is the experimental group and which the control is really not worth considering.

Example 1. A junior high school language arts teacher is in the process of designing a writing curriculum for seventh and eighth grades. Realizing that motivation is a strong determiner of junior high performance in any topic, the teacher has come up with four ways to motivate students to write; but of course he doesn't know which will work best, or if some will work better with some students than others. To give him this needed information for future planning, he has decided to perform a formative evaluation using one of the different strategies in each of his four roughly comparable, heterogeneously-grouped classes: One group will edit its own magazine; another will write articles to be submitted to popular national magazines; a third group will write letters to the editor of the local newspaper; and a fourth group will write a play about the problems of adolescence. The teacher hopes to take further advantage of this instance of Design 3, the Non-Equivalent Control Group Design (Chapter 4), by analyzing results on periodic writing exams separately for students whom he assessed to be good or poor writers in the first place.

Example 2. A district-wide Early Childhood Education program has decided to incorporate a psycho-motor development component that will require installation of large playground equipment. In order to answer many questions about the best way to integrate the program into the overall early childhood curriculum, the district's Assistant Superintendent for Early Childhood Programming has decided to install the equipment in two week phases to groups of randomly chosen schools. The entire pool of the district's elementary schools will be divided randomly into eight groups. The groups will receive the equipment and begin the program at two-week intervals. Having eight groups begin using the materials in this step-wise fashion will give the staff a chance to do formative evaluation. After administering a pretest, they will work with Group 1 for two weeks and then administer a psychomotor unit test. They will make necessary program modifications, then initiate the program with Group 2. Group 2's pretest results, because of randomization, should match Group 1's. Results of the unit test with Group 2, however, can be *compared* with results from Group 1 to determine if program modifications have had an effect on student development. This revision/program installation/test cycle can repeat as many as six more times, or until the program seems to be yielding maximal gains. This useful formative design is actually a version of Design 1, the True Control Group Pretest-Posttest, detailed in Chapter 4.

Tempered by proper caution about the danger of basing extremely important decisions on studies with small numbers, "formative planned variations" allow you to rest program planning on more than hunches.

2. *You might relax some of the more stringent requirements for implementing a design.* Since formative evaluation ~~usually~~ often collects information for the sole use of program staff, the formative evaluator can, where necessary, relax some of the requirements for setting up a design. This means that, when necessary, you can use assignment of students that is slightly less than random, or choose a non-randomized control group from students of a somewhat different socioeconomic group, *as long as interpretation of results is accompanied by appropriate caution.* The formative evaluator can at times relax design constraints because the formative evaluator's constituency is the program staff. They will use the data he gathers to make program change decisions. They will, in addition, serve not only as judges of what constitutes credible information but they will, through constant contact with the program, gather much of their own data—at least that concerned with attitudes and impressions. In situations when the formative evaluator has been able to set up comparison trials of various program versions, staff members inevitably gather first-hand experiences to use as a basis for making program revisions.

Regarding design, the job of the formative evaluator seems to be to provide many opportunities for comparison, using as good a design as possible. The details of the implementation of any one design are not critical.

Example. Jackson Elementary School, in the heart of a large urban area, received Federal funds to design a compensatory education program for the middle grades, with a particular focus on basic skills. The school identified the students eligible for the program according to the state's requirements for receiving the funds. By and large, these students were chronically low achievers. A young and devoted school staff had ideas about how best to use the money: they installed an Enrichment Center based on open school guidelines, and used much of the money to hire classroom aides. They were interested in keeping close watch on the quality of achievement that their first year of program operation produced, but they could not locate a control group. Someone suggested that the students in the school who traditionally performed slightly below average but not as poorly as the target students might form a rough control group for the study. Subsequently the decision was made that these students would be tested for progress in reading, math, and writing at the same times and using the same pre and post measures as the program students. This is a modification of Design 3, the Non-Equivalent Control Group design (Chapter 4), with a special awareness that the control group is indeed non-equivalent. The control group, for one thing, did score just significantly higher than the program students on a standardized pretest. A careful watch over the course of the school year, however, showed that program students received extensively more attention in basic skills areas and at the end of the year were achieving about the *same* as the control group. Such a design helped the staff conclude that the new program indeed did benefit the target students: they were now achieving as well as students who had scored better than them in the past.

An exception to this pronouncement about formative evaluation and more relaxed designs occurs in the case of controversies within the staff over different versions of program implementation. One of the jobs of the formative evaluator is to collect information relevant to differences of opinion about how the program should be designed or implemented. In this case, as with summative evaluation, challenges to the conclusiveness of results can occur, and credibility will become again important. Disagreements among planners can be translated into alternative treatments to form bases for small experiments designed according to the guidelines in this book.

3. *You might want to perform short experiments or pilot tests.* You will find that program planners must constantly make decisions about how a program will look. Most of these decisions must be made in the absence of knowledge about what works best. *Should all math instruction take place in one session, or should there be two during the day? How much discussion in the vocational education course should precede field trips? How much should follow? Will reading practice on the Readalot machine produce results as good as when children tutor one another? How much worksheet work can be included in the French*

course without damaging students' chances of attaining high conversational fluency? You can settle these questions by believing whoever offers the most convincing opinion, or you can subject them to a test. Using one of the evaluation designs described in this book, particularly Designs 1, 2, or 3, you can conduct a short study to resolve the issue. Read Chapters 2 and 3. Then choose treatments to be given to students (or whomever) that represent the decision alternatives in question. The duration of the short study should last as long as you feel will realistically allow the alternatives to show effects. If you will be reading this book for the purpose of designing short experiments, please substitute the word *treatment* for *program* as you read the text. The designs described in the book, and the procedures outlined for accomplishing them are, of course, equally appropriate.

Example. A group of third grade teachers attending a convention heard about a mathematics game which they thought would teach multiplication tables painlessly if played every Friday morning. Interested in saving their students the agony of drill, the teachers urged their principal to purchase the game. The principal, a former math teacher, was skeptical of the value of what she called "playing bingo." She refused. The teachers, however, persuaded the principal to agree to a test: they would randomly distribute students among their four classrooms every Friday morning for four weeks, carefully controlling the number of high and low math ability students distributed to each classroom. Two of the teachers would play the math game; the other two would drill their students in the same multiplication tables, and give prizes for knowing tables exactly like those to be won playing the game. At the end of the month, the data would be allowed to speak for themselves. This highly credible Design 1 study would uncover differences between drill and the program if any were to be attained.

Use of design requires planning in advance, if only to locate a group that is willing to serve as the comparison. Even if you have no intention of collecting comparative data at the outset, it might be a good idea to locate a handy group from whom you will be able to pull students in order to try out new lessons or plans or to do short experiments. Often you will find a teacher who is not taking part in the program who will be glad to provide you with a little time to give supplementary instruction, a short quiz or a questionnaire to his class.

Evaluation Where Design Presents Problems:
PROGRAMS AIMED AT SPECIAL POPULATIONS

Many evaluators find themselves in the position of collecting information about the quality of funded programs aimed at helping students, clients or others who are exremely rich or poor in a certain disposition, ability or attitude. In a school setting, for example, these special categories of children might score, for instance, in the top 2% on an IQ test and be labeled gifted, or below 75 IQ and be classified as retarded. The students may be handicapped or emotionally disturbed. Programs aimed at these students present unique design problems because laws requiring that all such children be educated rule out evaluation designs where the control group receives *no* special program. A comparison group can therefore only be formed if the school has *two* programs available for special students.

Example. A school tried two different kinds of programs for its gifted students. Gifted students were randomly assigned to one or another program for a 10-week trial period, at the end of which benefits from both programs were assessed by the principal. Reactions of students and parents were positive for both programs, but one program involving field trips raised considerable resentment from students not in the program. Since they were unable to justify the field trips as necessary, the principal and staff chose to continue with the other program.

The following paragraphs suggest other possible approaches to evaluation of special programs. The reader is also referred to How To Conduct Qualitative Studies for a discussion of alternative approaches.

1. *Use the non-equivalent control group design* (Design 3, Chapter 4). Such a comparison could be made if another district or school with no special programs, or programs appreciably different from yours, agreed to give the same tests as yours and to share results.

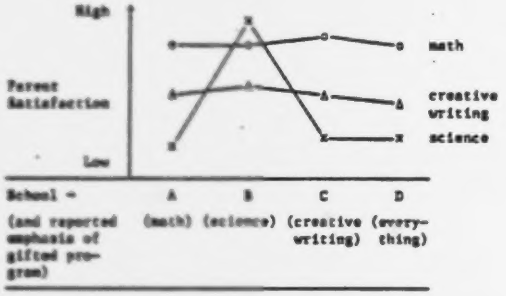
Example. Teachers of educable mentally retarded students planned a reading skills program which they hoped would significantly improve the reading of their EMR students. They asked a nearby elementary school to share with them results of a reading test given by the district in May each year and to permit a criterion-referenced test to be given to the EMR students at the beginning and end of the school year. Progress of the two groups in reading could be compared.

Adopt a formative approach and evaluate program components.

2. Comparative studies of the effects of whole programs are not always the best service you can provide to the program staff or even the funding agency. Rather, more useful information can be gained by evaluating *components* of a special education program with a view to recommending changes that might be needed in these. In some cases, for example, alternative materials might be available for teaching the same objectives. Small scale experiments could be set up in several schools, using a pretest-posttest true control group design (Design 1, Chapter 4) in each classroom to obtain objective data on the effectiveness of the various alternatives.

3. Compare diverse programs in terms of some common indicator, e.g., satisfaction with program outcomes using Design 3. Sometimes an evaluator is asked to evaluate a number of special programs which individual schools or projects have produced and which all have different goals and objectives.

For example, perhaps at one school, the gifted program concentrates on acceleration in math, at another on breadth of exposure in science, at another on creative writing, and at a fourth on all these things at once. You could measure at all schools student and parent satisfaction with the instruction provided in individual subjects (math, science, writing skills). Perhaps you would find results like this:



In general, parents seem equally satisfied with both math and creative writing, no matter what the emphasis reported by the school. Satisfaction with science, however, seems very sensitive to whether or not science is emphasized by the gifted program. When it is, there is high satisfaction. The evaluator might note that in the absence of special effort, science might not be well taught to gifted students, at least if parent satisfaction is a valid indicator.

The point of this example is that diverse programs can be assessed if you can find a single dimension on which to compare them. Opinions and attitudes often provide this common ground. This kind of investigation at least tells you what kind of programs seem to make a difference on the dimension you have chosen.

4. Compare program outcomes to pre-established criteria and use Design 6 (Before-and-After Design, Chapter 6). Frequently, special programs are required to state measurable goals, and the evaluator's job is to measure goal achievement. This often turns into a game of who can set goals which are lofty enough to be acceptable but simple enough to be reached, especially when goals are set in terms of standardized test gains. Sometimes, however, when the

goals are derived from criteria which have intrinsic, recognizable value, reasonable goal setting is an excellent approach. For example, specification of some basic survival skills, such as reading road signs correctly and making change, for retarded students, could provide mastery goals for an EMR program. A fairly good assessment of program effectiveness can be made even in the absence of a good design, if program results can be compared to reasonable goals.

5. *Make the evaluation theory-based.* A good approach to assessing the results of special ← programs is to do a *theory-based evaluation.* This is an evaluation that focuses on program *implementation,* holding the staff accountable for operating the program they have promised. The theory-based evaluation first asks: *On what theory of instruction, theory of learning, psychological theory, or philosophical point-of-view is the program based?* In other words, what activities does the staff view as critical to obtaining good results toward which the program aims? Detailed questioning of the staff makes explicit the model, theory, or philosophy that the staff is trying to implement. Once you know the staff's intention, your job will be to ascertain if activities that are specified by the theory are being effectively operationalized and implemented. Of course if you decide to do a theory-based evaluation, the existence of planned activities must be documented through objectively collected evidence, not just through testimonials. If you can show in your evaluation that the elements which the theory specifies as necessary for goal attainment are present, then you have shown that the program has taken an effective step toward goal achievement. If the theory is correct, goals should be reached eventually.

For Further Reading

Anderson, S. B. (Ed.). *New directions in program evaluation.* San Francisco: Jossey-Bass, Inc., 1978.

House, E. R. (Ed.). *School evaluation: The politics and process.* Berkeley: McCutchan, 1973.

Morris, L. L., & Fitz-Gibbon, C. T. Evaluator's handbook. In L. L. Morris (Ed.), *Program evaluation kit.* Beverly Hills: Sage Publications, 1978.

Popham, W. J. *Educational evaluation.* Englewood Cliffs, NJ: Prentice-Hall, 1975.

Struening, E. L., & Guttentag, M. *Handbook of evaluation research,* Vol. 1. Beverly Hills: Sage Publications, 1975.

Worthen, B. R., & Sanders, J. R. *Educational evaluation: Theory and practice.* Worthington, Ohio: Charles A. Jones Publishing, 1973.

HOW TO MEASURE PROGRAM IMPLEMENTATION

(DRAFT)

Revision Author: Jean King

December 6, 1985

Table of Contents

Chapter 1. Measuring Program Implementation: An Overview
Chapter 2. Questions to Consider in an Implementation Evaluation
Chapter 3. How to Plan for Measuring Program Implementation
Chapter 4. Methods for Measuring Program Implementation: Records
Chapter 5. Methods for Measuring Program Implementation: Self-Reports
Chapter 6. Methods for Measuring Program Implementation: Observations

Appendix A An Outline for an Implementation Report

Chapter 1

MEASURING PROGRAM IMPLEMENTATION: AN OVERVIEW

How to Measure Program Implementation is one component of the Program Evaluation Kit, a set of guidebooks written primarily for people who have been assigned the role of program evaluator. The evaluator has the often challenging job of scrutinizing and describing programs so that people may judge the program's quality as it stands or determine ways to make it better. Evaluation almost always demands gathering and analyzing information about a program's status and sharing it in one form or another with program planners, staff, or funders.

This book deals with the task of describing a program's implementation--i.e., how the program looks in operation.[1] Keeping track of what the program looks like in actual practice is one of the program evaluator's major responsibilities because you cannot evaluate something well without first describing what that something is. If you have taken on an evaluation project, therefore, you will need to produce a description of the program that is sufficiently detailed to enable those who will use the evaluation results to act wisely. This description may or may not be written. Even if delivered informally, however, it should highlight the program's most important characteristics, including a description of the context in which the program exists--its setting and participants--as well as its distinguishing activities and materials. The implementation report may also include varying amounts of backup data to support the accuracy of the description.

The overall objective of this book is to help you develop skills in describing program implementation and in designing and using appropriate instruments to generate data to support your description. The guidelines in the book derive from three sources: the experience of evaluators at the Center for the Study of Evaluation, University of California, Los Angeles; advice from experts in the fields of educational measurement and evaluation; and comments of people in school, system, and state settings who used a field test edition of the book. How To Measure Program Implementation has three specific purposes:

1. To help you decide how much effort to spend on describing program implementation
2. To list program features and activities you might describe in a program implementation report
3. To guide you in designing instruments to produce supporting data so that you can assure yourself and your audience that your description is accurate

The book has six chapters. Chapter 1 discusses the reasons for examining a program's implementation. Chapter 2 provides a list of questions that might be answered by an implementation evaluation. Chapters 3 through 6 comprise the "How to Measure" section of the book. Chapter 3 discusses how to plan an implementation evaluation, followed by three methods chapters devoted to an examination of existing records, self-report measures (questionnaires and interviews), and observation techniques.

Wherever possible, procedures in the "how to" sections are presented step-by-step to give you maximum practical advice with minimum theoretical interference. Many of the recommended procedures, however, are methods for measuring program implementation under ideal circumstances. It is no surprise that few evaluation situations in the real world match the ideal, and, because of this, the goal of the evaluator should be to provide the best information possible. You should not expect, therefore, to duplicate step-by-step the suggestions in this book. What you can do is to examine the principles and examples provided and then adapt them to your situation, whatever the evaluation constraints, data requirements, and report needs. This means gathering the most credible information allowable in your circumstances and presenting the conclusions so as to make them most useful to each evaluation audience.[2]

Why Look at Program Implementation?

One essential function of every evaluation is answering the question, "Does the combination of materials, activities, and administrative arrangements that comprise this program seem to lead to its achieving its objectives?" In the course of an evaluation, evaluators appropriately devote time and energy to measuring the attitudes and achievement of program participants. Such a focus reflects a decision to judge program effectiveness by looking at outcomes and asking such questions as the following: What results did that program produce? How well did the participants do? Was there community support for what went on in the program? Every evaluation should consider such questions.

But to consider only questions of program outcomes may limit the usefulness of an evaluation. Suppose evaluation data suggest emphatically that the program was a success. "It worked!" you might say. Unless you have taken care, however, to describe the details of the program's operations, you may be unable to answer the question that logically follows your judgment of program success, the question that asks, "What worked?" If you cannot answer that question, you will have wasted effort measuring the outcomes of events that cannot be described and must therefore remain a mystery. Unless the programmatic black box is opened and its activities made explicit, the evaluation may be unable to suggest appropriate changes.

If this should happen to you, you will not be alone. As a matter of fact, you will be in good company. Few evaluation reports pay enough attention to describing the program processes that helped participants achieve measurable outcomes. Some reports assume, for example, that mentioning the title and the funding source of the project provides a sufficient description of program events. Other reports devote pages to tables of data (e.g., Types of Students Participating or Teachers Receiving In-service Training by Subject Matter Area) on the assumption that these data will adequately describe the program's processes for the reader. Some reports may provide a short, but inadequate description of the program's major features (e.g., materials developed or purchased, teacher and student in-class activities, employment of aides, administrative supports, or provisions for special training). After reading the description the reader may still be left with only a vague notion of how often or for what duration particular activities occurred or how program features combined to affect daily life at the program sites.

To compound the problem of omitted or insufficient description, evaluation reports seldom tell where and how information about program implementation was obtained. If the information came from the most typical sources--the project proposal or conversations with project personnel--then the report should describe the efforts made to determine whether the program described in the proposal or during conversations matched the program that actually occurred. Few evaluations give a clear picture of what the program that took place actually looked like and, among those few that do provide a picture of the program, most do not give enough attention to verifying that the picture is an accurate one.

It could be argued that this lack of attention to detail and accuracy is justifiable in situations where no one wants to know about the exact features of the program. This, however, is a bogus argument because you simply cannot interpret a program's results without knowing the details of its implementation. For one thing, an evaluation that ignores implementation will add together results from sites where the program was conscientiously installed with those from places that might have decided, "Let's not and say we did." If achievement or attitude results from the overall evaluation are discouraging, then what's to be done? This scenario typifies a poor evaluation study, but unfortunately, it describes many large-scale program evaluations from the '70s, including a few of those most notorious for showing "no effect" in expensive Federal programs (e.g., the 1970 evaluation of Project Follow-Through).

What is more, ignoring implementation--even when a thorough program description is not explicitly required--means that information has been lost. This information, if properly collected, interpreted, and presented, could provide audiences now and in the future a picture of what good or poor education looks like. One important function of evaluation reports is to serve as program records. Without such documentation educators may continue to repeat the mistakes of the past.

Why look at program implementation? Two things should be clear by now:

> - Description, in as much detail as possible, of the materials, activities, and administrative arrangements that characterize a particular program is an essential part of its evaluation; and
>
> - An adequate description of a program includes supporting data from different sources to insure thoroughness and accuracy.

How much attention _you_ choose to give to implementation in your own situation, then, will substantially affect the quality of your evaluation. A detailed implementation report, intended for people unfamiliar with the program, should include attention to program characteristics and supporting data as described in Table 1.

[Insert Table 1]

What and How Much To Describe?

A quick look in Chapter 2 at the list of possible questions for an implementation evaluation will show you that assembling information and writing a detailed implementation report about even a small program could be an impossible job for one person who must work within the constraints of time and a budget. To help you in such a situation, the remainder of the chapter poses some questions to focus your thinking about what to look at, measure, and report. Considering these questions before you make decisions about measuring implementation should help insure that you spend the right amount of time and effort describing the program and use the measures most appropriate to your circumstances.

Before planning data collection about program implementation, you will need to make two decisions:

1. Which features of the program is it most critical or valuable for me to describe? This may amount to deciding which questions in Chapter 2 to use. Your answer will depend, in part, on how much time and money you have. It will also be affected by your role vis-a-vis the staff and the funding agency, the announced major components of the program, and the amount of variation allowed by its planners.

2. How much and what kind of data will be necessary to support the accuracy of the description of each program characteristic? Decisions about backup evidence will determine whether your report simply announces the existence of a program feature or offers evidence to support the description you have written. This decision will also be constrained by time and money, as well as by your own judgments about the need for corroboration and the amount of variation you have found in the program.

If you feel that your experience with evaluation or with the program, the staff, or the funding agency is sufficient to allow you to make these decisions right now, then process to Chapter 2 and being planning your data collection.

If you do not yet feel ready, the four questions that follow will give you further guidance toward making decisions about what to look at and how to back up your report. These questions relate to the following issues: (1) deciding whether you need to document the program or work for its improvement; (2) determining the most critical features of the program you are evaluating; (3) finding out how much variation there is in the program; and (4) deciding how much and what type of supporting data is needed.

Question 1. What Purposes Will Your Implementation Study Serve?

This question asks you to consider your role with regard to the program. Your role is primarily determined by the use to which the implementation information you supply will be put. The question of use will override any other you might ask about program implementation.

If you have responsibility for producing a summary statement about the general effectiveness of the program, then you will probably report to a funding agency, a government office, or some other representative of the program's constituency. You may be expected to describe the program, to produce a statement concerning the achievement of its intended goals, to note unanticipated outcomes, and possibly to make comparisons with an alternative program. If these tasks resemble the features of your job, you have been asked to assume the role of summative evaluator.

On the other hand, your evaluation task may characterize you as a helper and advisor to the program planners and developers. During the early stages of the program's operations, you may be called on to describe and monitor program activities, to test periodically for progress in achievement or attitude change, to look for potential problems, and to identify areas where the program needs improvement. You may or may not be required to produce a formal report at the end of your activities. In this situation, you are a trouble-shooter and a problem solver, a person whose overall task is not well-defined. If these more loosely-defined tasks resemble the features of your job, you are a formative evaluator. Sometimes an evaluator is asked to assume both roles simultaneously--a difficult and hectic assignment, but one that is usually doable.

While concerns of both the formative and summative evaluator focus on collecting information and reporting to appropriate groups, the measurement and description of program implementation within each evaluation role varies greatly, so greatly that different names are used to characterize the two kinds of implementaiton focus. Description of program implementaiton for summative evaluation is often called program documentation. A documentation of a program is its official description outlining the fixed critical features of the program as well as diverse variations that might have been allowed. Documentation connotes something well-defined and solid. Documentation of a program, its summative evaluation, should occur only after the program has had sufficient time to correct problems and function smoothly.

On the other hand, description of program implementation for formative evaluation can be called program monitoring or evaluation for program improvement. Monitoring connotes something more active and less fixed than documentation. The more fluid connotation of monitoring reflects the evolving nature of the program and its formative evaluation requirements. The formative evaluator's job is not only to describe the program, but also to keep vigilant watch over its development and to call the attention of the program staff to what is happening. Program monitoring in formative evaluation snould reveal to what extent the program as implemented matches what its planners intended and should provide a basis for deciding whether parts of the program ought to be improved, replaced, or augmented. Formative evaluation occurs while the program is still developing and can be modified on the basis of evaluation findings.

Measuring implementation for program documentation

Part of the task of the summative evaluator is to record, for external distribution, an official description of what the program looked like in operation. This program documentation may be used for the following purposes:

1. Accountability. Sometimes the expected outcomes of a program, such as heightened independence or creativity among learners, are intangible and difficult to measure. At other times program outcomes may be remote and occur at some time in the future, after the program has concluded and its participants have moved on. This kind of outcome, concerned, for instance, with such matters as responsible citizenship, success on the job, or reduced recidivism, cannot be achieved by the participants during the program. Rather, the program is intended to move its participants toward achievement of the objective. In such instances, where judging the program completely on the basis of outcomes might be impractical or even unfair, program evaluation can focus primarily on implementation. Program staff can be held accountable for at least providing materials and producing activities that should help people progress toward future goals. Alternative school programs, retraining programs within a company, programs

responding to desegregation mandates, and other programs involving shifts of personnel or students are examples of cases where evaluation might well focus principally on implementation. Though these programs might result in remote or fuzzy learning outcomes, the nature of their proper implementation can often be precisely specified.

¶Of course, you might need to measure implementation for accountability purposes in any case. Even when a program's objectives are immediate and can be readily measured, it is likely that the staff will be accountable for some amount of implementation of intended program features. They will need to show, in other words, where the money has gone. This role of program documentation has been called the signal function, a sign of compliance to an external agency, a report that says "We did everything we said we were going to." While some may belittle this type of evaluation, its successful and timely completion is often critical to continued funding, and its importance should not be underestimated.

2. Providing a lasting description of the program. The summative evaluator's written report may be the only description of the program remaining after it ends. This report should therefore provide an accurate account of the program and include sufficient detail so that it can serve as a basis for planning by those who may want to reinstate the program in some revised form or at another site. Such future audiences of your report need to know the characteristics of the site and the sorts of activities and

materials that probably brought about the program's outcomes.

3. Providing a list of the possible causes of the program's effects. While such cases are unusual, a summative evaluation that uses a highly credible design and valid outcome measures constitutes a research study. It can serve as a test of the hypothesis that the particular set of activities and materials incorporated in the program produces good achievement and attitudes. Here the summative report about a particular program has something to say to policy makers about programs using similar processes or aiming toward the same goals. The activities and materials described in the evaluator's documentation, in this case, are the independent or manipulated variables in an educational experiment.

   The development of evaluation thinking over the past twenty years has led away from the notion that the quantitative research study is the only and ideal form for an evaluation to take. In cases where variables cannot be easily controlled or where creating a control group will deprive individuals of needed services or training, evaluators should neither lament their fate nor demean the project. But in those few cases where an evaluator has the opportunity to design and conduct a research study in the traditional sense, the opportunity should not be wasted.

Knowing the uses to which your documentation will be put helps you to determine how much effort to invest in it. Implementation information collected for the purpose of accountability should focus on producing the required "signals" by examining those activities, administrative changes, or materials that are either specifically required by the program funders or have been put forward by the program's planners as major means for producing its beneficial effects.

The amount of detail with which you describe these characteristics will depend, in turn, on how precisely planners or funders have specified what should take place. If planners, for example, have prescribed only that a program should use the XYZ Reading Series, measuring implementation will require examining the extent of use of this series. If, on the other hand, it is planned that certain portions of the series be used with children having, say, problems with reading comprehension, then describing implementation will require that you look at which portions are being used, and with whom. You will probably need to look at test scores to insure that the proper students are using XYZ. The program might further specify that teachers working in XYZ with problem readers carry out a daily 10-minute drill, rhythmically reading aloud, in a group, a paragraph from the XYZ story for the week. If the program has been planned this specifically, then your program description will probably need to attend to these details as well. As a matter of fact, attention to specific behaviors is a good idea when describing any program where you see certain behavior occurring routinely. Program descriptions at the level of teacher and student behavior help readers to visualize what students have experiences, giving them a good chance to think about what it is that has helped the students to learn.

If accountability is the major reason for your summative evaluation then you must provide data to show whether--and to what extent--the program's most important events actually did occur. The more skeptical your audience, the greater the necessity for providing formal backup data. Concerns about the skeptical audience are elaborated in later questions in this chapter. [MAKE SURE THEY ARE.]

If you need to provide a permanent record of program implementation for the purpose of its eventual replication of expansion, try to cover as many as possible of the program characteristics listed in Chapter 2. The level of detail with which you describe each program feature should equal or exceed the specificity of the program plan, at least when describing the features that the staff considers most crucial to producing program effects. If additional practices typical of the program should come to your attention while conducting your evaluation, you should include these. You will need to use sufficient backup data so that neither you nor your audience doubt the accuracy or generality of your description.

When describing implementation for the purposes of accountability and leaving a lasting record of a program, the data you collect can be fairly informal, depending on your audience's willingness to believe you. You might talk with staff members, peruse school records, drop in on class sessions, or quote from the program proposal.

In cases where the reason for measuring implementation involves research or where there is potential for controversy about your data and conclusions, you will need to back up your description of the program through systematic measurement, such as coded observations by trained raters, examination of program records, structured interviews, or questionnaires. Carefully planned and executed measurement will allow you to be reasonably certain that the information you report trruly describes the situation at hand. It is important that the evaluator produce formal measures in cases where he himself wants to verify the accuracy of his program description. It is essential that he measure if he thinks he will need to defend his description of the program, that is, if he might confront a skeptic. An example from a common situation should illustrate this.

[INSERT EXAMPLE HERE]

Measuring implementation for program improvement

As has been mentioned, the task of the formative evaluator is typically more varied than that of the summative evaluator. Formative evaluation involves not only the critical activities of examining and reporting student progress and monitoring implementation; it also often measn assuming a role in the program's planning, development, and refinement. The formative evaluator's repsonsbilities specifically related to program implementation usually include the following:

1. *Insuring, throughout program development, that the program's official description is kept up-to-date, reflecting how the program is actually being conducted*. While for small-scale programs, this description could be unwritten and agreed upon by the few active staff members, most programs should be described in a written outline that is periodically updated. An outline of program processes written before implemetation is usually called a program plan. Recording what has taken place during the program's implementation produces one or more formative implementation reports. The task of providing formative implementation

reports--and often insuring the existence of a coherent program plan as well--falls to the formative evaluator.

¶ The topics discussed in the formative report could coincide with the headings in the implementation report outline in the Appendix. The amount of detail in which each aspect of the program is described should match the level of detail of the program plan. In many situations, the formative evaluator finds his first task to be clarification of the program plan. After all, if he is to help the staff improve the program as it develops, he and they need to have a clear idea at the outset of how it is supposed to look. If you plan to work as a formative evaluator, do not be surprised to find that the staff has only a vague planning document. Unless the program relies heavily on commercially published materials with accompanying procedural guides, or the program planners are experienced curriculum developers, planners have probably taken a wait-and-see attitude about many of the program's critical features. This attitude need not be bothersome; as long as it does not mask hidden disagreements among staff members about how to proceed, or cover up uncertainty about the program's objectives, a tentative attitude toward the program can be healthy. It allows the program to take on the form that will work best.

¶ It gives you, however, the job of recording what does happen so that when and if summative evaluation takes place, it will focus on a realistic depiction of the program. An accurate portrayal of the program will also be useful to those who plan to adopt, adapt, or expand the program in the future. The role of the evaluator as program historian or recorder is an essential one, as it is often the case that staff people simply have no time for such luxuries. Even as simple a record as notes from meetings, arranged chronologically, can provide helpful information at a later date.

2. Helping the staff and planners to change and add to the program as it develops. In many instances the formative evaluator will become involved in program planning--or at least in designing changes in the program as it assumes cleaner form. How involved she becomes will depend on the situation. If a program has been planned in considerable detail, and if planners are experienced and well versed in the program's subject matter, then they may want the formative evaluator only to provide information about whether the program is deviating from the program plan.

   ¶On the other hand, if planners are inexperienced or if the program was not planned in great detail in the first place, then the evaluator becomes an investigative reporter. Her first job might be to find out what is happening--to see what is going well and badly in the program. She will need to examine the program's activities independent of guidance from the plan, and then help eliminate weaknesses and expand on the program's good points. If this case fits your situation, use the list of implementation characteristics in Chapter 2 as a set of suggestions about what to look for or adopt the naturalistic approach described later.

   The formative evaluator's service to a staff that wants to change and improve its program could result in diverse activities. Two of them are particularly important:

a. The formative evaluator could provide information that prompts the staff and planners to reflect periodically on whether the program that is evolving is the one they want to have. This is necessary because programs installed at a particular site practically never look as they did on paper--or as they did when in operation elsewhere. At the same time, staff and planners will be persuaded to reexamine their initial thinking about why the processes they have chosen to implement will lead to attaining their objectives. Careful examination of a program's rationale, handled with sensitivity to the program's setting, could turn out to be the greatest service of a formative evaluator. The planners should have in mind a sensible notion of cause and effect relating the desired outcomes to the program-as-envisioned. Insofar as the program-as-implemented and the outcomes observed fail to match expectations, the program's rationale may have to be revised.

b. Controversies over alternative ways to implement the program might lead the formative evalautor to conduct small-scale pilot studies, attitude surveys, or experiments with newly-developed program materials and activities. Program planners, after all, must constantly make decisions about how the program will look. These decisions are usually based only on hunches about what will work best or will be accepted most readily. For instance: Should all math instruction take place in one session, or should there be two sessions during *the day? How much discussion in the vocational education course should precede field trips? How much should follow? Will practice on the Controlled Reading Machine produce results that are as good as those obtained when children tutor one another? How much additional paperwork will busy instructors tolerate? How much worksheet activity can be included in the French course without detracting from students' chances of attaining high conversational fluency?* These are good and reasonable questions that can be answers by means of quick opinion surveys or short experiments, using the methods described in most texts on the topic of research design.[3] A short experiment will require that you select experimental and control groups, and then choose treatments to be given to these groups that represent the decison alternatives in question. These short studies should last long enough to allow the alternatives to show effects. The advantage of performing short experiments will quickly become apparent to you; they provide credible evidence about the effectiveness of alternative program components or practices. At the same time, it must be remembered that the real world environment surrounding most evaluations makes even simple experiments difficult to conduct.

When measuring implementation for program improvement, the form of evaluation reports can and should vary greatly. Informal conversations with an influential staff member may have more effect than a typewritten report, and particularly a report loaded with statistical tables. Periodic meetings to discuss program problems and issues may update administrators and teachers, forcing them to think about the activities in which they are engaged far better than even a short written document could. One wellknown evaluator has gone so far as to have program personnel place bets on the likely outcomes of data analysis so they will have a vested interest in the results.

Whether you work as a summative or a formative evaluator, you will need to decide how much of your implementation report can rely on anecdotal or conversational information and still be credible, and how much your report needs to be backed up by data produced by formal or systematic measurement of program implementation. If what you describe can make a difference to those who might use it for any of the purposes mentioned, then your implementation report deserves all the time and effort you can afford.

Question 2. What Are the Program's Most Critical Characteristics?

Having determined the purposes--formative or summative (or both)--that your implementation study will serve, your identification of the program's critical features will help you further to determine two things:

- The specific questions your evaluation will address
- The level of detail you should use in describing the program

Three features common to all programs can form an initial outline of a program's critical characteristics: context; activities; and "theory." You can begin to describe the program by outlining the elements of the program's context--the tangible features of the program and its setting:

- The classrooms, schools, districts, or sites where the program has been installed
- The program staff--including administrators, teachers, aides, parent volunteers, secretaries, and other staff
- The resources used--including materials constructed or purchased, and equipment, particularly that purchased especially for the program
- The students or participants--including the particular characteristics that made them eligible for the program, their number, and their level of competence at the beginning of the program

These context features constitute the bare bones of the program and must be included in any summary report. Listing them does not require much data gathering on your part, since they are not the sort of data that you expect anyone to challenge or view with skepticism. Unless you have doubts about the delivery of materials, or you think that the wrong staff members or students may be participating, there is little need for backup data to support your description.

Another part of the context you would do well to consider is not tangible, but may be essential to understanding program functioning. This is the political context into which the program is set. It includes, for example, understanding what interest groups or powerful individuals are involved in the program, how funding was initially secured, the role of top managers, problems encountered in the program, and so forth. In some settings, none of this will matter; in others, such information will allow you to target your evaluation on what can be usefully addressed. While such information is unlikely to appear in formal evaluation documents, only a naive evaluator operates without an awareness of the political context, and he does so at his and his evaluation's risk.

In addition to context features, the second area to describe in looking for critical characteristics is that of program activities. Describing important activities demands formulating and answering questions about how the program was implemented, for example:

- What were the materials used? Were they used as intended?
- What procedures were prescribed for the instructors to follow in their teaching and other interactions with students? Were these procedures followed?
- In what activities were the participants in the program supposed to participate? Did they?
- What activities were prescribed for other participants--aides, parents, tutors? Did they engage in them?
- What administrative arrangements did the program include? What lines of authority were to be used for making important decisions? What changes occurred in these arrangements or lines of authority?

Listing the salient activites intended to occur in the program will, of course, take you much less time than verifying that they have occurred, and in the form intended. Unlike materials, which usually stay put and whose presence can be checked at practically any time, program activites may be inaccessible once they have occurred if they were not consciously observed or recorded. Counting them or merely noting their presence is therefore no small task. In addition, activities are more difficult to recognize than context features. Math games, microcomputers, aides, and science materials from Company X are easily identified; but what exactly does the act of reinforcement or acceptance of a student's cultural background look like when it is taking place?

Occurrence of intangible activities such as reinforcement or cultural acceptance cannot be simply observed and reported like an inventory of materials or a headcount of students. Even if they could be directly observed, you could not possibly describe all of them. You will have to choose which activities to attend to. Your choice of these activities will in large measure depend upon what your audience has said it needs to know in order to make informed decisions.

Once context and activities are delineated, the third and often the most difficult program feature to determine is what can be called, for want of a better term, the program's "theory." Every program, no matter how small, operates with some notion of cause and effect, that is, with a theory. Examples are numerous: If teenage parents learn parenting skills, their children will eat more nutritiously; if bilingual students receive reinforcement in their native language, their cognitive skills and self-concept will develop normally; if longterm employees undergo technical education, their job productivity will increase. Some programs (e.g., Montessori schools, E.S.T., or Camp Hill Villages) are systematically designed to implement the tenets of an explicitly stated model, theory, or philosophy. Others evolve their own theories, combining common sense, practice, and theoretical tenets from a variety of sources. The job for the evaluator is to discover this theory in order to better understand how the program is supposed to work and its critical characteristics in the eyes of program planners and staff.

On paper it sounds easy to describe a program's context, key activities, and "theory," but when you try to do it, critical details may prove elusive.. Three sources of information should help you decide what your evaluation should examine:

1. The program proposal or plan
2. Opinions of program personnel, experts, and yourself, based on assumptions about what makes an educational program work
3. Your own observations

Picking out critical program features from the plan or proposal

Some program proposals will come right out and list the program's most important features, perhaps even explaining why planners think these materials and activities will bring about the desired outcomes. But many will not, although if you look carefully, you may find clues about what is considered important. For instance, most proposals or documents describing a program will refer over and over to certain key activities that should occur. As a rule of thumb, the more frequently an activity is cited, the more critical someone considers it to be for program success. You may therefore decide that activities repeatedly mentioned are critical program components to which the evaluation must attend.

The program's budget is another index to its crucial features. As another rule of thumb, you may assume that the larger the budgeted dollar or other resource expenditure, such as staffing level, for a particular program feature--activity, event, material, or configuration of program elements--the greater its presumed contribution to program success. Taken together, these two planning elements--frequency of ciration and level of expenditure or effort--can provide some indication of the program's most critical components.

Relying on the program plan for suggestions about *what* needs to be described determines a point of view from which to approach your implementation evaluation. *An implementation evaluation based largely on the program plan will involve collecting data to determine the extent to which the crucial activities named in the plan occurred as intended* – and if they did not occur as planned, what happened instead. Description of a program from this point of view is the kind most often done, and for an understandable reason: it provides the simplest means by which the evaluator can decide which activities to look at.

Example. A group of health and science teachers wrote a proposal to the state for a few thousand dollars to assemble a personal hygiene and sex education course for Rooms City's high schools. The program was to be based largely on purchased audio-visual materials. The state evaluator who examined the program relied heavily on the original proposal as a program descriptor. To complete the documentation section of his summative report, he simply noted the program's official description, and observed informally to locate consistencies and discrepancies between the planned program and the one that actually occurred.

Even in the absence of a formal written program plan, documentation from the perspective of implicit planning can be done by *interviewing program planners* and asking them to describe activities they feel are crucial to the program. You can then proceed with the documentation of the extent of occurrence of these activities.

You might find the program plan and even the planners themselves, in some instances, to be disappointing sources of ideas about what to look for. They might not describe proposed activities to the degree of specificity you feel you need; or they might express grandiose plans, engendered by initial enthusiasm, or in response to proposal guidelines from the funding source which were themselves overly ambitious. It is possible, as well, that the program *has not been planned* in any specific way.

How then will you document the program if, for whatever reason, there is no plan which details activities that are specific, feasible, and consistent throughout? In this case, you have two options: you can rely on what *theory* and *experienced people* say should be in the program; or you can take the point of view of a ~~responsive~~ *naturalistic observer** and simply watch the program operating to discover what seems to be the program's critical features.

Relying on opinions of program personnel, experts, and yourself to select critical program features

If you have reason to believe that some feature *not mentioned* in the program's planning documents might be necessary for program success, then look for it. Commonly unmentioned but critical program characteristics, for instance, are *rehearsal or repetition of learned information*, ~~and~~ *adequate time on task*, and teacher training. Planners of instructional programs, it seems, spend a lot of time deciding what to teach and in what sequence, but often overlook students' need to repeat and study the information they have received, and teachers' need to alter their instruction.

Whether or not the kinds of characteristics or program activities mentioned above are critical in your situation, *you should give consideration to features, not specifically cited in the program plan, whose presence or absence might be related to program success or failure*. If you are a formative evaluator, it is, in fact, your responsibility to bring these matters to the attention of the staff. You might, incidentally, discover a feature of the program that someone thinks could actually make it fail. By all means, pay attention to this kind of information, backing up your description with data.

Example. Mr. Walker, the director and *de facto* formative evaluator of in-service training programs at a university-based Teacher Center, noticed that some districts sending teachers allowed them free choice of courses. Others, believing that in-service training should follow a theme, encouraged teachers to take courses within a single area say elementary math, or affective education.

Though the Teacher Center itself made no recommendations about what courses should be pursued, Mr. Walker decided that the "theme versus no-theme" flavor of teacher training might have an effect on teachers' overall assessment of the value of their in-service experiences. He decided to describe the course of study of the two groups of teachers at the Center and separately analyze the groups' responses to an attitude questionnaire.

As Mr. Walker expected, teachers whose training followed a theme expressed greater enthusiasm about the Teacher Center. Since he could find no explanation for the difference in enthusiasm between the two groups other than the thematic character of one group's program, Mr. Walker recommended that the Center itself encourage thematic in-service study. He used his descriptions of the courses of study of the teachers in the theme group as a set of models the Center might follow.

To the extent that you base your choice of what to look for on a set of assumptions about what works in education, you are conducting what could be called a "theory-based" evaluation.

Mr. Walker, in the example above, worked from the rather rudimentary but verifiable theory that education that follows a program of study is more likely to be perceived by the student as valuable. His evaluation was at least partly theory-based because he used a theory to tell him what to look at.

Examining program implementation in theory-based evaluation gives your study a point of view toward the program similar to the one you assume when basing implementation measurement on the proposal: you begin with a *prescription* of what effective program activities might look like. The prescription from the theory based perspective, however, comes not from a written plan, but from a theory.

A theory-based implementation evaluation is especially appropriate for looking at a school program that is built on a *model* of teaching behavior; *theory* of learning, development, or human behavior; or *philosophy* concerning children, schools, or organizations. The specific prescriptions of many such models and theories are familiar to most people working in education.[4]

Examples of some of these models are:

- Behavior modification and various applications of reinforcement theory to instruction and classroom discipline
- Piaget's theory of cognitive development and other models of how children learn concepts
- Open-classroom and free-school models such as those put forth by writers in education in the 1960's
- Fundamental-school competency based and basic skills models ~~which~~ that seek to reinstate traditional American classroom practices
- Models of organizations that prescribe arrangements and procedures for effective management

- Approaches to teaching critical thinking skills that encourage the use of higher level questioning techniques

A program identified with any of these points of view must set up roles and procedures consistent with the particular theory or value system. Proponents of open schools, for instance, would agree that a classroom reflecting their point of view should display freedom of movement, individualization of instruction, and curricular choices made by students. Each theory, philosophy, or teaching model contends that *particular activities* are either worthwhile in and of themselves or are the best way to promote certain desirable outcomes. Measuring implementation of a theory-based program, then, becomes a matter of checking the extent to which activities or organizational arrangements at the program sites reflect the theory.

Theories underlying programs may be intuitive and specific, as in Mr. Walker's example, or explicit and general.

Cooley and Lohnes have proposed a *general model* of school learning[5] that seems particularly useful as a source of ideas for what to look at when describing a program intended to *teach* people something. The effectiveness of school programs in bringing about desired learning, according to this model, depends on four factors:

1. *Learning opportunities.* Schools provide the time and place in which students may practice new skills, attend to sources of new information, or come in contact with models of how to act.
2. *Motivation.* Schools intentionally manipulate rewards and punishments that persuade students to pursue prescribed activities and attend to particular information.
3. *Structured presentation of activities, ideas, and information.* Schools attempt to organize and sequence what is presented, tailoring it to students' abilities so as to make learning as painless and efficient as possible.
4. *Instructional events.* The school day is filled with social and interpersonal contacts that promote learning. The elimination of misunderstandings through dialogue, a teacher's effective use of student contributions in a class discussion, and the personal attention and reassurance that prevent student discouragement are some examples of instructional events in this sense.

Figure 1 shows a simple diagram of the Cooley/Lohnes model. Opportunity, motivators, structure, and instructional events change the student from his level of initial performance to the criterion performance desired for the program. The important thing about the Cooley/Lohnes model for describing program implementation is that each of these four aspects of schooling involves critical features that an evaluator might want to mention when describing a program.

Some evaluators have shown that different philosophies about school and classroom processes can be described and compared nicely using the four dimensions of the Cooley/Lohnes model.[7] In looking at program implementation, an evaluator can ask questions such as the following:

- *Opportunity.* What materials addressing the program objectives were available to students? How much time was allotted to learning and practicing the target skills and understandings? How did conditions affecting opportunity, such as attendance, access to materials, and distractions from relevant learning activities, vary across sites or among students? Did learning opportunities vary *intentionally* in duration and nature? If so, were differences specified by the program, or left to teachers or students to determine?

- *Motivators:* Were the materials capable of maintaining student attention or interest? Was a reinforcement system used? What systems of reward or punishment were used, and were parents involved? Did success with motivation techniques vary across sites? What conditions might have had a bearing on the different levels of motivation among students?
- *Structure:* To what extent were program objectives specified? Were learning hierarchies used to underlie the curriculum? Was a coherent outline used? How much attention was paid to sequencing of lessons? Was there an in-service program to teach novel subject matter, and were teachers aware of the rationale behind program sequencing? What efforts were made to ensure that program objectives and instruction would be suitable in terms of students' backgrounds and abilities?
- *Instructional Events:* What interpersonal contacts were there that tended to support student involvement in program activities? How much personal attention did individual students receive? Was instruction one-to-one or one-to-many? Businesslike or friendly? Frequent or infrequent? Primarily between students and teachers, students and students, or students and aides?

Theory-based evaluation might also involve your assessment of the consistency of the program *plan* with the underlying theory.

In summative evaluations based on a credible research design, you should note, a theory-based evaluation can provide an actual *test* of the theory's validity. Given the potential importance, and rarity, of empirical validation of a theory, results of an evaluation which has provided such validation should be reported and disseminated as widely as possible.

Using observations and case studies to determine critical program features

The evaluation literature in recent years has been charged by a debate over the value of methods that have been variously called qualitative, naturalistic, ethnographic, responsive, or even "new paradigm." While each of these terms has its own proper definition, in common usage they together describe evaluation techniques borrowed largely from anthropology and sociology that generate words as products, rather than numbers. The skills they demand of an evaluator differ greatly from those required in the more traditional quantitative approach, and, while it is beyond the scope of this book to provide an indepth description of qualitative evaluation methods, reference and how-to's will be added where appropriate throughout the text. The use of observations and more detailed case studies to determine critical features of a program being evaluated is one such place and will necessarily involve qualitative methods.[8]

It is possible for an evaluator with a qualitative mindset to observe a program in operation with relatively few preconcpetions or decisions about what to look for. This strategy might be chosen for a number of reasons. For one thing, ~~some~~ many evaluators consider it the best way of describing a program. Unhampered by preconceptions and prescriptions, the responsive/naturalistic inquirer might set his sights on catching the true flavor of a program, discovering the unique set of elements that make it work, and conveying them to the evaluation's audience. Further, a naturalistic approach might be necessary if there is no written plan for the program you are evaluating and you find that one cannot be retrospectively constructed with a reasonable degree of consistency by the planners. Even if there is a plan, it might be vague or, from your perspective, unrealistic to implement. Then again, you might discover that the program has been allowed so much variation from site to site that common features are not apparent at first. In any of these cases, you have the option of just observing.

Implicit in your decision to use qualitative responsive methods are at least two other decisions:

1. To rely heavily on data collection methods that "get close to the data," usually ~~classroom~~ on-site observations and interviews
2. To concentrate on relating what you found, rather than comparing what was to what should have been.

This ~~might~~ may leave you up in the air at first about what to look for.

Example. The School Board of a small city decided that high schools should spend one year emphasizing Language Arts, with particular focus on improving students' writing skills. The district's Assistant Superintendent for Curriculum resisted the initial impulse to design and implement a common, districtwide program. Instead, she decided that each teacher should be allowed to respond, in his or her own way, to the basic decision to emphasize writing. Her reasoning was that some teachers would arrive at good methods that the other teachers could use to everyone's advantage--students and teachers alike. To keep track of what teachers were doing, however, she scheduled periodic teacher and student interviews and dropped in on class sessions frequently. She wrote vignettes describing classroom practices she had seen and which reflected the aspirations and reports of teachers and students. Her report demonstrated to the Board the effects of its priority decision; and circulated among teachers in abbreviated form, it served as a source of new teaching ideas.

This ~~scenario might look~~ approach may seem familiar to you. The evaluator's vignettes correspond to how most people share information; and indeed, a ~~responsive~~ qualitative evaluation in a context that is free from controversy or skepticism looks very much like what people usually do. The difference between this evaluation, however, and a formal ~~responsive~~ naturalistic evaluation is in the quality of the observations made. ~~Responsive~~ Qualitative evaluators use methods from the social sciences—notably anthropology—to obtain corroboration for their observations and conclusions. They have, in fact, developed a method for conducting evaluations that follows that of naturalistic field studies.

An evaluation using ~~a~~ ~~naturalistic~~ qualitative methods would follow a scenario something like this:

1. A particular program is to be evaluated. If there are numerous sites, one or more sites is chosen for study.

2. The evaluator observes activities at the site or sites chosen, perhaps even taking part in the activities, but trying to influence the program routine as little as possible. Often, time constraints require the use of "informants"—people who have already been observing things and who can be interviewed.
3. Though data collection could take the form of coded records like those produced through the standard observation methods described in Chapter 6, the ~~responsive/~~naturalistic observer more often records what he sees in the form of *field notes*. This choice of recording method is motivated mainly by a desire to avoid deciding too soon which aspects of the situation observed will be considered most important.
4. The ~~responsive/~~naturalistic observer shifts back and forth between formal data collection, study of recorded notes, and informal conversation with the subjects. Gradually she produces a description of the events and direct or indirect interpretation of them. The report is usually an oral or written narrative, though naturalistic studies yield tables, sociograms, and other numerical and graphic summaries as well.

Case studies (considered technically) represent not so much a *method* as a choice of *what* to study. Case study researchers quite often follow ~~naturalistic~~ qualitative methodology. The case study worker in evaluation chooses to examine closely a particular case—that is, a school, a ~~classroom~~ work unit, a particular group, or an individual experiencing the program. Sometimes the program itself is "the case." Whereas the naturalistic observer or the more traditional evaluator might concentrate only on those experiences of, say, a school, ~~which~~ that are related to the program, the case study evaluator will usually be interested in a broader range of events and relationships. If the school is the subject of study, then the job is to describe the school. The case study method places the program within the context of the *many* things which happen to the school, its staff, and its students over the course of the evaluation. One result of this method, you can see, is to display the proportional influence of the program among the myriad other factors influencing the actions and feelings of the people under study. While case studies often use naturalistic methods presumably because of the complexity of the experiences and encounters which need to be described it is possible for a case study to use more traditional methods of data collection as well as to subject the case to ~~a contrived situation~~ the more controlled methods ~~reminiscent~~ reminiscent of traditional experiments.

Regardless of how you determine the list of critical characteristics, ~~A~~ a listing of the critical features of the program will give you some notion of which questions in Chapter 2 to answer in your implementation ~~report.~~ evaluation. If you are a summative evaluator, then your task will be to convey to your audience as complete a depiction of the program's crucial characteristics as possible.

If you are a formative evaluator, then your decision about what to look at might have to go a step beyond listing the program's critical features. Since your job is to help with program improvement and not merely to describe the program, your task is to collect information that will be maximally useful *for helping the program staff to improve the program.* In most cases, this will certainly mean monitoring the implementation of the program's most critical features. But you will need to consult with the program staff to find which among all the program's critical features seem most troublesome to them, most in need of vigilant attention, or most amenable to change. It could be, for instance, that a program's most critical feature is employment of aides. But once the aides have arrived and it has been established that they come to work regularly, attention to this detail may not be necessary. Your formative service to the program will be more usefully employed in monitoring the implementation of program aspects about which the staff has genuine problems to solve.

Question 3. How Much Variation Is There in the Program?

Your choice of which program characteristics to describe will be influenced by the *amount of variation* that occurs across sites where the program is being used and variation that happens at different points in time. For one thing, depending on the point of view of the planners, variability might be considered desirable or undesirable. Some programs, after all, *encourage* variation. Directors of such programs have said to the staff or to their delegates at different sites something like the following:

> *The district curriculum office has chosen six reading programs which we can purchase with our new Federal Compensatory Education money. Examine these, and select the one you think best suits your students and teachers.*

It is likely that an evaluator either formative, or summative will be called in to examine the whole Federal Compensatory Education program, and he will probably find six versions of the program taking place. Here variation across sites *has been planned,* and implementation of each reading subprogram will have to be described separately. Where such planned variation occurs, incidentally, the evaluator has a good opportunity to collect information that might be useful for future planning in this district or elsewhere particularly if the district is forced ~~wants~~ to narrow the number of reading programs to fewer than six. He can compare the ease and accuracy of implementation and success with students of the various programs across sites. Where different programs have been implemented by sites that are otherwise similar, the evaluator can compare results to gain clues about the relative effectiveness of the programs.

Program directors could have allowed the program to vary in an even *less* controlled way by saying:

> *We have X dollars to improve our reading program for the educationally disadvantaged. Take these funds and put together a new program.*

This kind of directive produces a program whose only common features across sites are likely to be the target students and the funding source! While variation is also *planned* in this kind of situation, unlike the program in the preceding example, each site has been left free to create *its own unique program.* The district-wide evaluator will have to look separately at each different version of the program that emerges, probably adopting a theory-based, case study, or judgmental ~~responsive/naturalistic~~ method. Though he may find a chance to make comparisons among the program variations put into effect at each site, he will probably spend a great deal of time *discovering* and reporting about what each program variation looked like. However, the simple act of telling the implementors about the various forms the program has taken will be useful. Most probably, some forms of the program will be more easily implemented, produce better results, or be more popular than others.

A program can afford to permit considerable variation across sites only in its early stages when it can make mistakes with minimum fear of penalties. For this reason, dealing with planned variation should be primarily the concern of the formative evaluator whose responsibility would then entail tracking the variations, comparing results of different versions of the program at comparable sites, and sharing information about commendable practices. Unfortunately, funding agencies often request summative reports at a time in the life of a program when considerable variation still exists. When this happens, the summative evaluator should note that several *different* program renditions are being evaluated. He should describe each of these, *and report results separately*, making comparisons where possible.

If the evaluator—whether summative or formative—should uncover variation across sites or over time that has *not* been planned, then he will have to describe this, collecting backup data if he feels that he will need corroborating evidence.

Question 4. When Do You Need Supporting Data?

You might need to describe program implementation for people who are at some *distance* from the program, either in terms of location or familiarity. These people will base their opinions about the program's form and quality on what they read in your description. You might therefore need to provide backup data to verify its accuracy.

If the description you produce is for people *close* to the program and familiar with it, then you can rely on the audience's detailed knowledge of the program in operation--at least in their own setting. In such a case, you may want to focus your data collection on the extent to which the program's implementation at one site is representative of its implementation at other sites. The credibility of your report for people close to the program will, of course, depend on how well your description of the program matches what they see. If you feel that your report of overall program implementation diverges considerably from the experiences of the program's administration or of participants at any one site, then you may need to collect good, hard backup data.

Examples of more specific circumstances calling for backup data are:

- Summative evaluations which constitute research studies addressed to the educational community at large
- Evaluations aimed at providing new information for a situation where there is likely to be controversy
- Evaluations calling for program implementation descriptions so detailed that they characterize program activity at the level of teacher or student behaviors
- Descriptions of programs that may be used as a basis for adopting or adapting the program in other settings
- Descriptions of programs which have varied considerably from site to site or from time to time

How you use backup data will be determined in part by which of the approaches to describing the program you adopt:

1. Using the program plan as a baseline and examining how well the program as implemented fits the plan.
2. Using a theory or model to decide the features that should be present in the program. In this case you will probably consult research literature or prescriptions of various philosophical or psychological points of view for guidance in what to look for. In both this and the plan-based approaches, backup data will be necessary to permit people to judge how closely the actual program fits what was planned. Such data could also help you document your discovery of program features that were not planned.
3. Following no particular prescription and instead taking a responsive/naturalistic stance regarding the program. In this situation you will attempt to enter the program sites with no initial preconceptions or assumptions about what the program should look like.

If you assume either of the first two points of view concerning focus and use of data, your final report will describe the fit of the program to the prescription you have chosen to use. In the third situation, your final report will simply describe the program that you found, noting, of course, variability from site to site.

This chapter has discussed the measurement of program implementation with a view toward making this aspect of your evaluation report reflect the needs of your audiences, the context you are working in, and your own professional standards. To help ensure that your reports will be useful and credible, this chapter has been concerned with the critical decisions you should make *before* you begin your evaluation—which features of the program your evaluation should focus on and how you will substantiate your description of the program. To help you with these decisions, your attention has been directed to ~~three~~ four key questions:

1. What purposes will your implementation study serve?
2. What are the program's most critical characteristics?
3. How much variation is there in the program?
4. When do you need supporting data?

Your implementation evaluation should be as methodologically sound as you can make it. And, as when dealing with achievement and attitudes, your report should provide *credible* and, above all, *useful* information to your audiences.

Endnotes (These will become footnotes in the published version)

1. In general, describing program implementation is considered synonymous with measuring attainment of process objectives or determining achievement of means-goals, phrases used by other authors. The book prefers, however, not to discuss implementation solely in connection with process goals and objectives. This is because the primary reason for measuring implementation in many evaluations is to describe the program that is occurring--whether or not this matches what was planned. Other times, of course, measurement will be directed solely by pre-specified process goals. Describing program implementation is a broad enough term to cover both situations.

2. Audience is an important concept in evaluation. The audience is the evaluator's boss; she is its information gatherer. Unless she is writing a report that will not be read, every evaluator has at least one audience. Many evaluations have several. An audience is a person or group who needs the information from the evaluation for a distinct purpose. Administrators who want to keep track of program installation because they need to monitor the political climate constitute one potential audience. Curriculum developers who want data about how much achievement a particular program component is producing comprise another. Every audience needs different information; and, important, each maintains different criteria for what it will accept as believable information.

3. See, for instance, ??? UPDATED VERSION OF HOW TO DESIGN A PROGRAM EVALUATION. See, also, the "Step-by-Step Guide for conducting a small experiment" in ???, UPDATED VERSION OF EVALUATOR'S HANDBOOK.

4. An excellent presentation of the implications of various models of schooling and education is put forth in Joyce, B., & Weil, M. *Models of teaching.* Englewood Cliffs, NJ: Prentice-Hall, 1972. See, as well, Kohl, H. The open classroom. New York: Random House, 1969 and also Neill, A. S. *Summerhill.* New York: Hart, 1960.

5. Cooley, W. W., & Lohnes, P. R. *Evaluation research in education.* New York: Irvington Publishers, 1976.

6 ~~4~~. Reprinted from Cooley, W. W., & Lohnes, P. R. *Evaluation research in education.* New York: Irvington Publishers, 1976, p. 191.

7. Leinhardt, G. Applying a classroom process model to instructional evaluation. *Curriculum Inquiry,* 1978, 8(2).

8 For a more detailed discussion of how to conduct a qualitative evaluation, see Patton, Michael Q., (Kit Book on Qualitative Methods) and other references listed at the end of this chapter

9. Where there is *one* formative evaluator working with the program district-wide, she will become involved with assessing variation and perhaps sharing ideas across sites. Where there is a separate formative evaluator *at each site,* each evaluator will work according to different priorities. The job of each evaluator will be to see that *each* version of the program develops as well as possible, perhaps disregarding what other sites are doing.

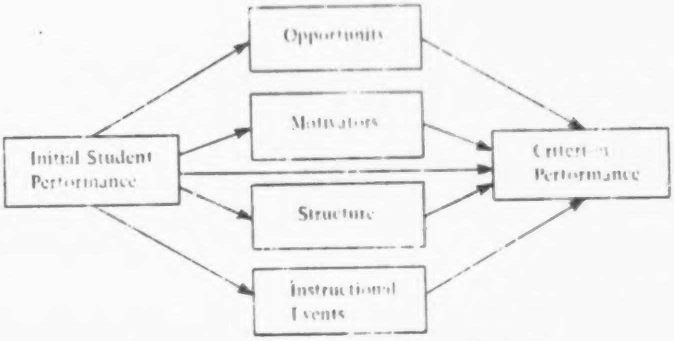

Opportunity
Motivators
Initial Student Performance
Criterion Performance
Structure
Instructional Events

Figure 1. A model of classroom processes

Chapter 2

QUESTIONS TO CONSIDER IN PLANNING AN IMPLEMENTATION EVALUATION
ONE HUNDRED QUESTIONS FOR AN IMPLEMENTATION EVALUATION

Chapter 1 presented an introduction to implementation evaluation, including a rationale for conducting both formative and summative evaluations and four questions to help you ~~initiate~~ begin ~~the~~ planning ~~process for~~ such an evaluation. The purpose of this chapter is to help you continue your initial planning by listing many things you might want to know about a program you're evaluating--but might not think to ask. Such a claim to inclusiveness is at least in part facetious; every program has unique features that will generate questions of a highly individual nature. But at the same time, the list of questions that follows, generated from the experiences of many evaluators with many programs, may help you to focus on aspects of the program you might not otherwise have seen as important.

It may help you to think of this list as an outline for what you may want to report to individuals who will use the results of your evaluation. The headings and questions in this chapter are organized according to what could eventually become the ~~five~~ major sections of a formal implementation report. But at this point in your evaluation, you should not worry about what your final product may look like. Research on evaluation use has taught us that useful reports take a variety of forms, from casual conversations over coffee, to working meetings, to formal documents typed and bound.

At this stage in your planning, you needn't worry about report format, but rather about the specific information you need to collect in order to answer the program's most important questions. If you are conducting a formative evaluation for immediate program improvement, jot down questions that would enable you to quickly provide information for suggesting strengths or for effecting changes. If, on the other hand, yours will be a summative evaluation documenting a program's implementation, target instead questions that will enable you to create a meaningful record of what happened in or as a result of the program.

In addition to choosing _what_ to describe about the program, you will need to decide which portions of your description must be supported by corroborating evidence. The necessity to collect supporting data to underlie your description of some program features will, of course, be primarily a function of the setting of _your_ evaluation. But there are program features which, because of their complexity, controversial nature, or critical weight within programs, usually require backup data regardless of the context. To remind you that your description of certain features may meet skepticism, an asterisk appears in the outline next to questions whose answers could require accompanying evidence.

Outline of remainder of chapter-- The list of questions will be divided into three sections, as follows:

A. Program overview
    1. Setting (what is the program's context?)
    2. Program origins and history

   3. Rationale, goals, objectives
   4. Program staff and participants
   5. Administration and budget

B. Program specifics (critical characteristics of the program)
   1. Planned program characteristics
   2. Questions for examining program materials
   3. Questions for examining program activities

C. The evaluation itself
   1. Purpose and focus
   2. Range of measures and data collection
   3. Timeframe

Summary section at the chapter end emphasizing that not all questions fit every evaluation and that you won't always write up every bit of information you get (i.e., use these questions to help you frame a good evaluation, focus the evaluation process on the issues where you can or should make a difference).

Chapter 3

HOW TO PLAN FOR MEASURING PROGRAM IMPLEMENTATION

Chapter 1 listed reasons for including an accurate program description in your evaluation report. These reasons included the need to set down a concrete description of the program that could be used for its replication, to provide a basis for making conjectures about relationships between implementation and program effects, and to collect accountability evidence demonstrating that the program staff delivered the service they promised. The *summative* evaluator will be concerned about documenting a program's implementation for one or more of these purposes. The *formative* evaluator, on the other hand, will primarily be concerned about tracking changes in a program's implementation, keeping a record of the program's developmental history, and giving feedback to the program staff about bugs, flaws, and successes in the process of program installation.

Looking through Chapter 2 probably helped you to decide which characteristics of the program need describing. At some point in your thinking about program implementation, you will make a related decision about which of these descriptions need substantiating: that is, which parts of your report need to be backed up by data which you collect.

If you are a summative evaluator, the simplest way to describe program activities, materials, and administration but unfortunately the least adequate for most purposes is to use an existing description of the program (i.e., the plan or proposal) to double as your program implementation report. If you are severely pressed by time and other constraints, and if a plan exists, you *may* get by with this. But if you or a member of your staff have time to spend on actually *measuring* implementation, then your description of the program will be richer and subsequently more useful and credible. If you are a formative evaluator whose job is to report about what is going on at the program sites, you cannot help but become involved in some implementation measurements. For those who will do the measuring, this chapter presents several methods that might help you obtain backup data for your report.

Methods of Data Collection

This book introduces you to three common approaches for collecting backup data for your implementation report. The use of any one method does not exclude use of the others. Your selection of data collection methods depends on the extent to which, given available resources, your report will provide information that your audience considers accurate and credible. The method or combination of methods you select will be primarily a function of three factors: the overall purpose of your evaluation; the information needs of your audience; and the practical constraints surrounding the evaluation process.

Method 1. Examine the Records Kept Over the Course of the Program. The first method of data collection requires that you examine the program's existing records.

> These might include sign-in sheets for materials, library loan records, individual student assignment cards, teachers' logs of activities in the classroom. In a program where extensive records are kept as a matter of course, you may be able to extract from them a substantial part of the data you need to determine what activities occurred, what materials were used, and how and with whom activities took place and materials were used. This method will yield credible evaluation information because it provides evidence of program events accumulated *as they occurred* rather than reconstructed later. The major drawback of existing records is that abstracting information from them can be time consuming. ~~Then again~~, records kept over the course of the program will probably not meet all your data collection requirements. If it looks as though the existing records are inadequate, you have two alternatives. The best one is to set up *your own record-keeping system*, assuming, of course, that you have arrived on the scene in time to do this. A weaker alternative is to gather *recollected* versions of program records from participants. Should you do this, point out in your report the extent to which this information has been corroborated by more formal records or results from other measures.

Method 2. Use Self-Report Measures. A second data collection method involves having program personnel and participants--teachers, aides, parents, administrators, and students--provide descriptions of what program activities look like.

3-3

It makes sense, of course, to turn for information about a program to the people who worked with it. You might choose to *interview* people or give them *questionnaires*. If collecting information from *everyone* who experienced the program will take too much effort and time, then ask for descriptions of activities from *a sample* of people within each role group.

Since different groups of participants in a program might have divergent perceptions, you may want to gather self-report information, probably on a sample basis, from teachers, administrators, parents, and students and compare the information provided by different groups to see if you get a consistent, ~~or reliable~~, set of pictures about the program.

Be aware that *self-report* measures ~~could~~ have credibility problems depending on the situation. Usually people close to the program will find such information to be credible. People far from the program, for instance, at the funding agency, are less likely to trust self-report information from the staff. First of all, there is the possibility that people providing you with information have a vested interest in making the program look good. Then ~~again~~ Second, even when intentional bias is unlikely, self-report descriptions of a program are at best second-hand accounts of what transpired: the *evaluator* tells the *audience* what people *say they did*. Thirdly, self-report information often consists of recollections after-the-fact of peoples' *own* behavior. Accounts of what people remember having done themselves are usually not as credible as descriptions by others who actually *saw* what they did.

Because of their credibility problems and the detail with which program implementation usually needs to be described, self-report instruments are more often used to verify or to check on the *consistency* across sites of a program description arrived at by more direct means. Only when the evaluator's resources are too limited to permit collection of close-up data do self-report measures constitute the primary source of implementation information.

Method 3. Conduct Observations. The final data collection method discussed here is that of actual observations of program activities, having one or more observers make periodic visits to program sites to record their observations, either freely or according to a pre-determined list of questions. Although it can require a great deal of time and effort, on-site observation has high credibilty because the observer watches or even participates in program events as they occur. You can enhance that credibilty by demonstrating that the data from the observations are reliable, i.e., consistent across different observers and over time.

To help you in thinking about which methods of data collection are appropriate for your own situation, Table 2, pages ?? and ??, summarizes the advantages and disadvantages of each of the methods. The remainder of the book then devotes a chapter to presenting each of the three data collection methods in greater detail:

> Chapter 4 discusses how to use records to assess program implementation, describing both how to check program records that already exist and how to set up a record-keeping system.
>
> Chapter 5 describes ways of using self-report instruments with staff members, parents, students, etc., giving step-by-step procedures for constructing and administering questionnaires and conducting interviews.
>
> Chapter 6 discusses program observations, describing methods for conducting both informal and systematic observations. The section on systematic observation presents several alternative schemes for coding information, one of which should fit your needs.

If it is important that you describe a program feature accurately and/or if your audience might be skeptical, then you should try to assemble *converging* data. This requires using *multiple measures and data collection methods* and gathering data from different participants at different sites. For example, if you were evaluating a program based on individualization, you might want to document the extent to which instruction really is determined according to individual needs. To assure enough evidence, you could collect different kinds of data. Maybe you would *interview* students at the various program sites about the sequence and pacing of their lessons and the extent to which instruction occurs in groups. To corroborate what you find through student interviews, you could examine the teachers' *record-keeping* systems. In an individualized program it is likely that teachers would maintain charts or prescription forms tracking individual student progress. Finally, you might conduct a few *observations* or spot checks, watching typical classes in session to estimate the amount of individual instruction and progress-monitoring per student both within and across sites. Three sources of information—interviews, examination of records, and classroom observation—could then be reported, each supporting or qualifying the findings of another.

3-5

Where To Look For Already Existing Measures

Before you involve yourself in the onerous business of designing your own implementation measures, you might take a look at instruments already available. Some measures, mainly observation schedules and questionnaires, have been developed which can be used to describe general characteristics of groups, classrooms, and other educational units. Titles of these instruments often mention:

- School or classroom climate
- Patterns of interaction and verbal communication
- Characteristics of the environment

If you wish to explore some of these, check the~~se~~ following anthologies:

Borich, G. D., & Madden, S. K. *Evaluating classroom instruction: A sourcebook of instruments.* Menlo Park, CA: Addison-Wesley Publishing, 1977.

This sourcebook contains a comprehensive review of instruments for evaluating instruction and describing classroom activities. It lists 171 instruments, describes each along with its availability, reliability, validity, norms, if any, and procedures for administration and scoring. Each is also briefly reviewed, and sample items are provided. Only measures which have been empirically validated appear in the sourcebook. The instruments are cross-classified according to what the instrument describes (teacher, pupil, or classroom) and who provides the information (the teacher, the pupil, an observer).

Boyer, E. G., Simon, A., & Karafin, G. R. (Eds.). *Measures of maturation.* Philadelphia: Research for Better Schools, Humanizing Learning Program.

This is a three-volume anthology of 73 early childhood observation systems. Most of these systems were developed for research purposes, but some can be used for program evaluation.

The 73 systems are classified according to:

- The kinds of behavior that can be observed (individual actions and social contacts of various types)
- The attributes of the physical environment
- The nature and uses of the data and the manner in which it is collected
- The appropriate age range and other characteristics of those observed

Each system is described in detail.

Simon, A., & Boyer, E. G. *Mirrors for behavior: An anthology of classroom observation instruments.* Philadelphia: Research for Better Schools, Center for the Study of Teaching, 1974.

This collection provides abstracts of 99 classroom observation systems. Each abstract contains information on the subjects of the observation, the

3-6

setting, the methods of collecting the data, the type of behavior that is recorded, and the ways in which the data can be used. In addition, an extensive bibliography directs the reader to further information on these systems and how they have been used by others.

An earlier edition of this work (1967) provides detailed descriptions of twenty-six of these systems.

Price, J. L. *Handbook of organizational measurement.* New York: Heath, 1972.

This handbook lists and classifies measures which describe various features of organizations. The measures are applicable, but not limited to, schools and school districts. The instruments are classified according to organizational characteristics, e.g., communication, complexity, innovation, centralization. The text defines each characteristic and its measurement. Then it describes and evaluates instruments relevant to the characteristic, mentioning validity and reliability data, sources from which the measure can be obtained, and references for additional reading.

Planning for Constructing Your Own Measures

Regardless of which methods you finally choose, your information gathering should include four important considerations, each of which should be thought through (outlined? addressed?) before you begin data collection. These planning bases are the following:

1. A list of the activities, materials, and administrative procedures on which you will focus
2. Consideration of the validity and reliability of the measures you will use
3. A sampling strategy, including a list of which sites you will examine, who will be contacted, interviewed, or observed, as well as when and how often
4. A plan for data summary and analysis

1. Constructing a list of program characteristics

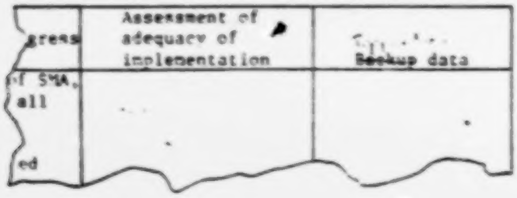
Composing a list of critical characteristics is the first step in each of the data gathering procedures outlined in Chapters 4, 5, and 6. Constructing an accurate list early in your evaluation will help insure that program decision makers receive credible information they will later be able to use.

A thoughtful look through the program's plan or proposal, a talk with staff and planners, your own thinking about what the program should look like—perhaps based on its underlying theory or philosophy—and careful consideration of the implementation questions in Chapter 2 should help you arrive at a *list of the program materials, activities or administrative procedures* whose implementation you want to track. Make sure that the program features you list are detailed and exhaustive of those considered—by the staff, planners, and other audiences—to be crucial to the program. *Detailed* means the list should include a prescription of the *frequency* or *duration* of activities and of their *form* (who, how, where) that is specific enough to allow you to picture each activity in your mind's eye.

If you are looking at a plan or proposal, then critical features will often be those *most frequently cited* and those to which the *largest part of the budget* and other resources have been allotted. For example, if large sets of curriculum materials were purchased for the program, then one critical part of the program implementation is the proper use of these materials.

If your work with the program will be *formative*, then you should attend to parts of the program that are likely to need revision or cause problems. Try to *visit* one or more sites in which the program is operating and observe the environment, the materials, the people, and the activities before you consider your list of program features complete. This way, you will be able to envision the actual program situation when you construct implementation instruments.

The program characteristics list can take any form that is useful to you. If you think you might use it later in a summative report, or as a vehicle for giving formative monitoring reports to staff, consider using a format like the one in Table 3, page 59. This table can serve as a standard against which to measure implementation. For summative evaluation, Table 3 could convey adequacy of implementation by adding two additional columns at the right:



gress	Assessment of adequacy of implementation	Backup data
of SMA, all ed		

You might prefer to begin with a less elaborate materials/activities/administrative features list than is shown in Table 3. The following example presents a simpler one.

3-7B

Example. The proposal for Emerson School's peer tutoring program contained the following paragraph: "... Tutoring activities will take place three days a week in the third, fourth, and fifth grade classrooms during the 45-minute reading period. Group 1 (fast) readers will each be assigned one slower reader whose reading seatwork will become their responsibility. All tutoring will be done using the exercises in the "Read and Say" workbooks which were purchased for the program. During tutoring, one teacher and one aide per classroom will circulate among student pairs, answering questions and informally monitoring the progress of tutees. Tutor-tutee rotation will take place every two months ..."

The assistant principal, given the job of monitoring the program's proper implementation, constructed for her own use a list of program characteristics which included her own informal notes:

Peer-Tutoring Activities

From written plan

- Frequency--3 times a week
- Duration--45-minute session
- Who--3rd, 4th, 5th graders
- Where--classrooms
- Fast readers teach slower
- Must "have responsibility"--what does this mean? (Director says it just means they will tutor same child all the time)
- All tutoring from "Read and Say"--in order, or can they skip around? (Third grade teacher says in order)
- Teacher and aide travel from pair to pair
- They "monitor"--is there a formal record-keeping system? (Director says yes--recording sheets have been drawn up and provided)
- Tutor-tutee rotate after two months

Additional data from interview with Bill Cox, Reading Specialist and Project Director, and Ms. Jones, third grade teacher:

- 3rd, 4th, and 5th grade tutors in their own classrooms; no switching rooms
- Teachers and aides--any difference in roles vis-à-vis tutors? No
- What did average readers do? Worked alone or in pairs with other average readers; tutored when a tutor was absent--does this cause disruptiveness?

2. Consideration of the validity and reliability of the measures you will use

Once you have constructed your list of program characteristics, you should next think in a general way about the type and content of instruments that would be appropriate for collecting data on those characteristics that are of most interest to your audience or those that have the potential for controversy. One important consideration in your planning is the technical adequacy of the implementation measures you will choose--the validity and reliability of methods used to assess program implementation. Even if you are not a statistical whiz, you should make sure that the instruments you eventually use will help you produce an accurate and complete description of the program you are evaluating.

> Assessments of the validity and reliability of a measurement instrument help to determine the amount of faith people should place in its results. *Validity* and *reliability* refer to different aspects of a measure's credibility. Judgments of validity answer the question:
>
> *Is the instrument appropriate for what needs to be measured?*
>
> Judgments of reliability answer the question:
>
> *Does the instrument yield consistent results?*
>
> These are questions you must ask about any method you select to back up your description of program implementation. "Valid" has the same root as "valor" and "value": it indicates how worthwhile a measure is likely to be for telling you what you need to know. Validity boils down to whether the instrument is giving you the true story, or at least something approximating the truth.
>
> When reliability is used to describe a measurement instrument, it carries the same meaning as when it is used to describe friends. A reliable friend is one on whom you can count to behave the *same way time and again.* In this sense, an observation instrument, questionnaire, or interview schedule that gives you essentially the same results when readministered in the same setting is a reliable instrument.
>
> But while reliability refers to *consistency*, consistency does not guarantee *truthfulness.* A friend, for instance, who compliments your taste in clothes *each time she sees you* is certainly reliable but may not necessarily be telling the truth. Further, she may not even be deliberately misleading you. Paying compliments may be a habit, or perhaps her judgment of how you dress may be positively influenced by other good qualities you possess. It may be that by a more objective standard you and your friend have terrible taste in clothes! Similarly, simply because an instrument is reliable does *not* mean that it is a good measure of what it seems to measure.

You are *measuring*, rather than simply *describing* the program on the basis of what someone *says* it looks like, because you want to be able to back up what you say. You are trying to assure both yourself and your audience that the description is an accurate representation of the program as it took place. You want your audience to accept your description as a substitute for having an omniscient view of the program. Such acceptance requires that you anticipate the potential arguments a skeptic might use to dismiss your results. When measuring program implementation, the most frequent argument made by someone skeptical of your description might go something like this:

> **Respondents to an implementation questionnaire or subjects of an observation have an idea of what the program is *supposed* to look like *regardless of whether this is what they usually do in fact*. Because they do not wish to appear to deviate, or because they fear reprisals, they will bend their responses or behavior to conform to a model of how they feel they *ought* to appear. Where this happens, the instrument, of course, will not measure the true implementation of the program. Such an instrument will be invalid.**

In measuring program implementation, concern over instrument validity boils down to a four-part question: Is the description of the program which the instrument presents *accurate*, *relevant*, *representative* and *complete*?[23]

An accurate instrument allows the evaluation audience to create for themselves a picture of a program that is close to what they would have gained had they actually seen the program. A relevant implementation measure calls attention to the *most critical features* of the program—those which are most likely related to the program's outcomes and which someone wishing to replicate the program would be most interested in knowing about.

A *representative* description of program implementation will present a *typical* depiction of the program and its sundry variations as they appeared across sites and over time. A *complete* picture of the program is one that includes *all* the relevant and important program features.

Making a case for accuracy and relevance

You can defend the *accuracy* of your depiction of the program by ruling out charges that there is purposeful bias or distortion in the information.

There are various ways to guard against such charges. Self-report instruments, for example, can be anonymous. If you are using observations, you can demonstrate that the observers have nothing to gain by a particular outcome and that the events they have witnessed were not contrived for their benefit. Records kept over the course of the program are particularly easy to defend on this account if they are complete and have been checked periodically against the program events they record. You need only show that the people extracting the information from the records are unbiased.

You can, in addition, show that administration procedures are *standardized*, that is, that the instrument has been used in the same way every time. Make sure that:

- Enough time was allowed to respondents, observers, or recorders so that the use of the instrument was not rushed
- Pressure to respond in a particular way was absent from the instrument's format and instructions, from the setting of its administration, and from the personal manner of the administrator

Another way to argue that your description is accurate is to show that results from any one of your instruments coincide logically with results from other implementation measures.

You can also add support to a case that your instrument is accurate by presenting evidence that it is *reliable.* Though it is usually difficult to demonstrate statistically that an implementation instrument is reliable, a good case for reliability can be based on the instrument's having *several items that examine each of the program's most critical features.* Measuring something important, say the amount of time students spend per day reading silently, by means of one item only, exposes your report to potential error from response formulation and interpretation. You can correct this by including several items whose results can be combined to compile an *index* ~~(see page 75)~~, or by administering the item several times to the same person.

If *experts* feel that a profile produced by an implementation instrument hits major features of the program or program component you intend to describe, then this is strong evidence that your data are *relevant.* For instance, a classroom description would need to include the curriculum used, the amount of time spent on instruction per unit per day, etc. A district-wide program, on the other hand, might need to focus heavily on key administrative arrangements for the program.

Making a case for representativeness and completeness

To demonstrate representativeness and completeness, you must show that in *administering* the instrument you ~~did~~ will not omit any sites or time periods in which program implementation may ~~have~~ look~~ed~~ different. You must also show that you have not given too much emphasis to a single atypical variation of the program. Thus your data must sample program sites typical of each of the different places where the program has been implemented. Your sample should also account for different times of the day, or different times during the life of the program if these are variations likely to be of concern. The variations you have been able to detect must represent the range of those that occurred.

As you can see, there is no one established method for determining validity. Any combination of the types of evidence described here can be used to support validity. If you plan to use an implementation instrument more than once, consider the whole period of its use an opportunity to collect information about the accuracy of the picture it gives you. Each administration is a chance to collect the opinions of experts, to assess the consistency of the view that this instrument gives you with that from other instruments, etc. Establishing instrument validity should be a continuing process.

¶ Reliability refers to the extent to which measurement results are free of unpredictable kinds of error. For example, if you were to give students a math test one day and, without additional instruction, give them the same test two days later, you would expect each student to receive more or less the same score. If this should turn out *not* to be the case, you would have to conclude that your instrument is *unreliable,* because, without instruction, a person's knowledge of math does not fluctuate much from day to day. If the score fluctuates, the problem must be with the test. Its results must be influenced by things other than math knowledge. These other things are called *error.*

Sources of error that affect the reliability of tests, questionnaires, interviews, etc., include:

- Fluctuations in the mood or alertness of respondents because of illness, fatigue, recent good or bad experiences, or other temporary differences among members of the group being measured.
- Variations in the conditions of use from one administration to the next. These range from various distractions, such as unusual outside noises, to inconsistencies and oversights in giving directions.
- Differences in scoring or interpreting results, chance differences in what an observer notices, and errors in computing scores.
- Random effects caused by examinees or respondents who guess or check off alternatives without trying to understand them.

BEST COPY

> Methods for demonstrating an instrument's reliability whether the instrument is long and intricate or composed of a single question usually involve comparing the results of one administration of the instrument with another by correlating[24] them.
>
> The evaluator designing and using instruments for measuring program implementation has unique problems when attempting to demonstrate reliability. Most of these problems stem from the fact that implementation instruments aim at characterizing a situation rather than measuring some quality of a person. While a person's skill, say in basic math, can be expected to stay constant long enough for assessment of test reliability to take place, a program cannot be expected to hold still so that it can be measured. Because the program will likely be dynamic rather than static, possibilities for test-retest and alternate form reliability are usually ruled out. And since most instruments used for measuring implementation are actually collections of single items which independently measure different things, the possibility of computing split-half reliabilities practically never occurs.

Few program evaluators have the luxury of sufficient time to design and validate data collection measures. But early attention to the validity and, to a lesser extent, the reliability of measures will help insure that the information gathered during the evaluation will enable the evaluator to answer well the questions that potential users most care about. An implementation evaluation can be a waste of time if it collects data that are technically "good," but that don't answer the right questions. Perhaps worse is the evaluation that relies on data that are weak at best. When decision-makers use bad data to guide program decisions, evaluation has done a disservice.

---

[24] Correlation refers to the strength of the relationship between two measures. A high *positive* correlation means that people scoring high on one measure also score high on the other. A low correlation means that knowing a person's score on one measure does not educate your guess about his score on the other. Correlations are usually expressed by a *correlation coefficient*, a decimal between -1 and +1, calculated from people's scores on the two measures. Since there are several different correlation coefficients, each depending on the types of instruments being used, discussion of how to perform correlations to determine validity or reliability is outside the scope of this book. The various correlation coefficients are discussed in most statistics texts, however. You might also refer to *How To Calculate Statistics*, part of the *Program Evaluation Kit*.

3. Creating a Sampling Strategy

Unless the program you are examining is short and simple, *you will not be able to collect and transcribe data on every student and activity over the course of the entire program.* What is more, there is no need to cover the entire spectrum of sites, participants, events, and activities in order to produce a complete and credible evaluation. But you will need to decide *early* where the implementation information you do collect will come from. Specifically you must plan:

- Where to look
- Whom to ask or observe
- When to look—and *how* to sample events and times

Where to look

The first decision concerns *how many* program sites you should examine. Your answer to this will be largely determined by your choice of measurement method; a questionnaire, for instance, can reach many more places than can an observer. Unless the program is taking place in just a few places, close together, it will probably not be practical or necessary to examine implementation at all of them. *A representative sample will provide you with sufficient information to be able to develop an accurate portrayal of the program.*

Solving the problem of *which* sites constitute a representative sample requires that you first group them according to two sets of characteristics:

1. Features of the *sites* that could affect how the program is implemented—such as size of the population served, geographical location, number of years participating in the program, amount of community or administrative support for the program, level of funding, teacher commitment to the program, student or staff traits or abilities
2. Variations permitted in the *program* itself that might make it look different at different locations such as amount of time given to the program per day or week, choice of curricular materials, or omission of some program components such as a management system or audiovisual materials.[15]

The list of such features is long and unique to each evaluation. For your own use, choose four or so likely sources of major program divergence across sites and classify the sites accordingly. Then, based on how many sites you think you can examine, try to randomly choose some to represent each classification. You can, of course, select some sites for intensive, perhaps even case, study and a pool of others to examine more cursorily.

You may also, for public relations reasons, need to at least make an appearance at every program site. In any case, make certain that you will be allowed access to every site you will need to visit. Such access should be assured before you begin collecting data.

---

[15] Where possible, including a few comparable sites which have not installed the program at all will give you a basis for interpreting some of the data you collect. This will help you determine, for instance, whether the absentee rate in the program is unusual or how much added effort is required from instructors. You can gather similar comparison data by monitoring or asking about usual practice at the program site before it was initiated.

Whom to ask or observe

Regardless of the size of the program or how many sites your implementation evaluation reaches, you will eventually have to talk with, question, or observe *people*. In most cases, these will be people both *within* the program—the participants whose behavior it directs—and those *outside*—parents, administrators, contributors to its *context*. Answers to questions about whether to *sample* people depend, as with your choice of sites, on the measurement method you will use and your time and resources.

Whom you approach for information also depends on the willingness of people to cooperate, since implementation evaluation nearly always intrudes on the program or consumes some staff time. If you plan to use questionnaires, short interviews, or observations that are either infrequent or of short duration, then you probably can select people randomly. In these cases, applying the clout factor by having a *person in authority* introduce you and explain your purpose will facilitate cooperation.

If you intend to administer questionnaires or interviews *for other purposes*, perhaps to measure people's attitudes, you may be able to insert a few implementation questions into these. It is often possible, and good practice, to consolidate instruments.

At times your measurement will require a good deal of cooperation. This is the case with requests for record-keeping systems that require continuous maintenance; intensive observation, either systematic or responsive/naturalistic; and questionnaires and interviews given periodically over time to the same people. If data collection requires considerable effort from the staff, and you have too little authority to back your requests, then you should probably ask for voluntary participants. Possible bias from volunteerism can be checked through short questionnaires to a random sample of other staff members. The advantage of gathering information from people willing to cooperate is that you will be able to report a complete picture of the program.

Exactly *which* people should you question or observe? Answers to this will vary, but here are some pointers:

- Ask people, of course, who are likely to know—key staff members and planners. If you think that these people might give you a distorted view, your audience will likely think so too. Thus you should back up what official spokespersons tell you by observing or asking others.
- Some of the *others* should be students if possible. Good information also comes from support staff members, assistants, aides, tutors, student teachers, secretaries, parents. People in these roles see at least part of the program in operation every day—but they are less likely to know what it is supposed to look like *officially*.
- Ask people to nominate the individuals who are in the best position to tell you the "truth" about the program. When the same names are mentioned by several program people, you know that you should carefully consider the information they provide.

If you intend to observe or talk to people several different times over the course of the program, then choice of respondents will be partially dependent on your time frame. Choosing which times and events to measure is discussed in the next section.

When to look

Time will be important to your sampling plan if your answer to any of these questions is *yes*:

- Does the program have phases or units that your implementation study needs to describe separately?
- Do you wish to look at the program periodically in order to monitor whether program implementation is on schedule?
- Do you intend to collect data from any individual site more than once?
- Do you have reason to believe that the program will change over the course of the evaluation?
- If so, do you want to write a profile of the program throughout its whole history that describes how it evolved or changed?

In these situations, you will probably have to sample *data collection dates.* First, divide the time span of the program into crucial *segments*, such as beginning, middle, and end; first week, eighth week, thirteenth week; or Work Units 1, 3, and 6. Then decide if you will request information from the *same* sample of people, at *each* time period or whether you will set up a *different* sample each time.

If and when you sample, be sure to return to the pool the sites or staff members selected to provide data during one particular time segment so that they might be chosen again during a subsequent time segment. People (or sites) should not be eliminated from the pool because they have already provided data. Only when you *sample* from the entire group can you claim that your information is representative of the entire group.

Timing of data collection needs additional adjustment for each measurement method. Questionnaires and interviews that ask about typical practice can be administered at any time during the period sampled. Some instruments, though, will make it necessary to carefully select or sample particular *occasions.* You want your *observations,* for instance, to record typical program events transpiring over the course of a typical program day. The records you collect should not come from a period when atypical factors—such as a bus strike or flu epidemic—are affecting the program or its participants. Sampling of specific occasions—days, weeks, or possibly even hours—will be necessary, as well, if you plan to distribute self-report measures which ask respondents to report about what they did "today" or at a specific time.

Figure 2 demonstrates how selection of sites, people, and times can be combined to produce a sampling plan for data collection. In Figure 2, a district office evaluator has selected *sites*, *people* (roles), and *times* in order to observe a reading program in session. The sampling method is useful because, in essence, the evaluator wants to "pull" representative *events* randomly from the ongoing life of the program. Her strategy is to construct an implementation description from short visits to each of the four schools taking part in the program.

Figure 2 is an example of an *ex*tensive sampling strategy; the evaluator chose to look *a little* at *a lot* of places. Sampling can be *in*tensive as well—it can look a lot at a few places or people. In such a situation, data from a few *sites*, classrooms, or students can be assumed to mirror that of the whole group. If the set of sites or students is relatively *homogeneous*, that is, alike in most characteristics that will affect how the program is implemented, you can randomly select representatives and collect as much data as possible from them exclusively. If the program will reach heterogeneous sites, classrooms, groups of students, etc., then you should select a representative sample from *each* category addressed by the program—for instance, schools in middle class versus schools in poorer areas; or fifth grades with delinquency-prone versus fifth grades with average students. Then examine data from each of these representatives. The strategy of looking intensively at a few places or people is almost always a good idea whether or not you use extensive sampling as well. These intensive studies could almost be called *case studies*, except that most case study methodologists disavow the need to ensure representativeness.

4. Planning Data Summary and Analysis

This section is intended to help you consolidate the data you collect regardless of your evaluation's purpose or intended outcomes.

There are two possible purposes for an implementation study. The first and major one is, of course, to *describe* the program and perhaps comment about how well it matches what was intended. A second purpose is to examine *relationships* between program characteristics and outcomes or among different aspects of the program's implementation. Examining relationships means exploring usually statistically—the hypothesis on which the program is based. *Do smaller classes achieve more? Are periodic planning meetings related to staff morale?*

It may seem odd to be concerned about how you will summarize the data at a point where you have barely decided what questions to ask. But it is time-consuming to extract information from a pile of implementation instruments and record, examine, summarize, and interpret it. Thinking about the data summary sheet in advance will encourage you to eliminate unnecessary questions and make sure you are seeking answers at the appropriate level of detail for your needs.

To handle data efficiently, you should prepare a *data summary sheet* for each measurement instrument you use—if possible, at the time you design the instrument. Data summary sheets will help you interpret the backup data you have collected and support your narrative presentation because they assist you in searching for *patterns of responses* that allow you to characterize the program. They also assist you in doing calculations with your data, should you need to do so.

The following section has four parts:

- A description of the use of *data summary sheets* for collecting together item-by-item results from questionnaires, interviews, or observation sheets, pages 67 to 71.
- Directions for reducing a large number of narrative documents, such as diaries/ or responses to open-ended questionnaires or interviews into a shorter but representative narrative form, pages 71 and 72.
- Directions for *categorizing* a large number of narrative documents so that they can be summarized in *quantitative* form, page 73.
- Suggestions for analyzing and reporting quantitative implementation data, pages 74 to 77.

Preparing a data summary sheet for scoring by hand or by computer

A data summary sheet requires that you have *either closed-response data or data that have been categorized and coded.* Closed-response data include item results from structured observation instruments, interviews, on questionnaires. These instruments produce tallies or numbers. If, on the other hand, you have item results that are narrative in form, as from open-ended questions on a questionnaire, interview, or naturalistic observation report, then you will *first* have to categorize and code these responses if you wish to use a data summary sheet. Suggestions for coding open-response data appear on page 73.

The first part of the following discussion on the use of summary sheets deals with recording and analyzing by hand; the latter part deals with summary sheets for machine scoring and computer analysis.

When scoring by hand, you can choose between two ways of summarizing the data: the quick-tally sheet and the people-item roster.

A *quick-tally sheet* displays all response options for each item so that the number of times each option was chosen can be tallied, as in the examples on page 68.

The quick-tally sheet allows you to calculate two descriptive statistics for each *group* whose answers are tallied: (1) *the number or percent of persons who answered each item a certain way*, and (2) *the average response to each item* (with standard deviation) in cases where an average is an appropriate summary. Notice that with a quick-tally sheet, you "lose" the individual person. That is, you no longer have access to individual response patterns. That is perfectly acceptable if all you want to know is how many (or what percentage of the total group) responded in a particular way.

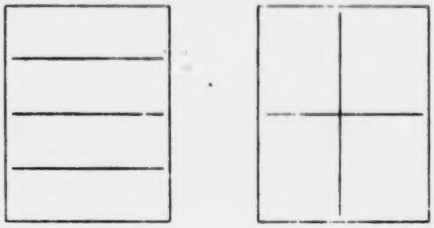
Often, for data summary reasons or to *calculate correlations*, you will need to know about the response patterns of *individuals* within the group. In these cases, a *people-item data roster* will preserve that information. On a people-item data roster the items are listed across the top of the page. The people (or classrooms, program sites, etc.) are listed in a vertical column on the left. They are usually identified by number. Graph paper, or the kind of paper used for computer programming, is useful for constructing these data rosters, even when the data are to be processed by hand rather than by computer. The people-item data roster below shows the results recorded from the filled-in classroom observation response form that precedes it.

INSERT XEROXED PORTIONS OF PAGES 70-71

How to Summarize a Large Number of Written Reports Into Shorter Narrative Form

If you have to summarize answers to open response questionnaire items, diary or journal entries, unstructured interviews, or narrative reports of any sort, you will want a systematic way to do this. The following list, adapted to your own needs, should help you to design your own system for analysis.

1. Begin by writing an identification number on each separate data source (e.g., each questionnaire, each journal). If used properly, these numbers will always enable you to return to the original if need be to check its exact wording.
2. If you have sufficient time, read quickly through the materials you are trying to summarize, looking for major themes, categories, and issues, as well as critical incidents and particularly expressive quotations. Mark all of these in pencil.
3. If you will analyze the data by hand, obtain several sheets of plain paper to use as tally sheets. Divide each paper into about four cells by drawing lines.



If you have access to a microcomputer and can type well enough, consider using it instead of paper so you will avoid having to copy data by hand.

3-19

4. Select one of the reports, and look for the kinds of events or situations it describes or, if you completed step 2, for evidence of any major categories you have already determined. As soon as an event is described, write a short summary of it in a cell on one of the tally sheets. You may also wish to copy an exact quotation if it is particularly well worded. Be sure to include the ID number of the report in parentheses following the summary so that if you should need to return to the original you will not have to go through all of the reports to find it. Then, in one corner of the cell, tally a "1" to indicate that that statement has been made in one report.

As you read the rest of the report, every time you come upon a previously unmentioned event, summarize it in a cell and give it a single tally for having appeared in one report. When you have read through the entire report, put a checkmark or other mark on it to indicate that you have finished with it. If you are summarizing open-ended questionnaire results and having 30 or fewer respondents, you might want to copy the responses to each item in order to put in one place the specific answers to a given question.

5. Read the rest of the reports in any order. Record *new* statements as above. When you come upon one that *seems to have been mentioned in a previous report,* find the cell that summarizes it. Read carefully, making sure that it is more or less the same kind of event. Record another "|" in the cell to show that it has been mentioned in another report. If some *part* of an event or opinion differs substantially from or adds a significant element to the first, write a statement that covers this different aspect in another cell so that you may tally the number of reports in which this new element appears.

6. Prepare summaries of the *most frequent* statements for inclusion in *your* report. There may be good reasons for recording separately data from different groups if the reporters faced circumstances that would predictably bring about different results (e.g., different grade levels, different program variations). Also, if the quantity of the data that you are gleaning from the reports appears to be unwieldy, you may find it necessary to organize the events mentioned into different categories--in some cases more general, in others more narrow. Whenever new summary categories are formed, however, you are cautioned to avoid the blunder of trying to transfer previous tallies from the original categories. The only safe procedure is to return to the original source, the reports themselves, and then tally results for the new categories.

How to summarize a large number of written reports by categorizing

The following procedure helps you to assign numerical values to different types of responses and use this data in further statistical analyses. Suppose, for example, you asked 100 teachers to describe their experiences at a Teacher Learning Center where they received in-service training in classroom management techniques. After reading their reports and summarizing them for reporting in paragraph form, you wonder how closely the practices of the Teacher Center conform to the official description of the instruction it offers. You can find this out by *categorizing* teachers' reports into, say, five degrees of closeness to official Teacher Center descriptions--very close, through so-so, to downright contradictory--giving each teacher an opinion score, 1 through 5. Such rank-order data will give you a quantitative summary of teachers' experiences of the program. Perhaps you could then correlate this with their liking for the program or their achievement in courses.

The difficulty of the task of categorizing open-response data will vary from one situation to another. Precise instructions for arriving at your categories and summarizing your data cannot be provided, but the following advice should help make the task more manageable:

1. Think of a *dimension* along which program implementation might vary–closeness of fit to the program plan, perhaps, or approximation to a theory, or effectiveness of instruction. The dimension you choose should characterize the kinds of reports given to you so that you can put them in order from desirable to undesirable.
2. Read what you consider to be a representative sampling of the data–about 25%. Determine if it is possible to begin with three general categories: (a) clearly desirable, (b) clearly undesirable, and (c) those in between.
3. If the data can be divided in these three piles, you can then put aside for the moment those in categories (a) and (b) and proceed to refine category (c) by dividing it into three piles:
    - Those that are more desirable than undesirable
    - Those that are more undesirable than desirable
    - Those in between
4. Refine categories (a) and (b) as you did (c). If you cannot divide them into three gradations along the dimension you have chosen, then use two; or if the initial breakdown seems as far as you can go, leave it as is.
5. Have one or more people check your categories. This can be done by asking others to go through a similar categorization process or to critique the categories and the selections you have made.

Some suggestions for analyzing and reporting quantitative implementation data

Computing results characteristic by characteristic. If you want to report *quantitative* information from your implementation instruments, this section is designed to help you. It assumes that you have first transferred data to a summary sheet.

Your implementation data depict the frequency, duration, or form of critical characteristics of the program. If you want to explore relationships between certain program characteristics and others, or between program features and achievement or attitude outcomes of the program, then you want to make statements on the nature of: "Programs which had characteristic K tended to J." Here K is a description of the frequency or form of a particular program feature, and J is an achievement, the attitude in a particular group, or perhaps the frequency or form of yet another program feature. You might, for instance, want to see whether programs with more than two aides in the classroom show higher staff morale; or perhaps whether experience-based high school vocational programs with a wide choice of work study plans have fewer dropouts.

Showing this relationship can be done in two ways:

- You can use instrument results (K) to classify programs and then calculate the average J per program, or
- You can *correlate* K with J.

Before you bother to compute a statistic, you should be clear about the question you are trying to answer, and consider who would be interested in the answer and what impact it might have.

If you decide to explore relationships of this sort, you have two choices about what to use for K (and J, if it is another program feature):

1. K can be a summary of responses to a *single item.* It could be, for instance, a classification of schools by funding level of the program, or the average number of participating classrooms at a site. It could be the number of parent volunteers, the number of years the program has been in operation, or observers' estimate of the average amount of time spent at a particular activity. If you use a single item to determine this classification, then make sure that the item gives valid and reliable information. The probability of making an error when answering one item is usually so large that people might be skeptical. If you must use a single item to indicate K, then make sure you can verify what the item tells you. If the classification according to program characteristics which gives you K is critical to the evaluation, you should probably use multiple measures or an *index* to estimate K.
2. You can calculate an *index* to represent K by combining the results of several items or several different implementation measures. A procedure that asks about slightly different aspects of the same characteristic several times, and then combines the results of these questions to indicate the presence, absence, or form of the characteristic, is less likely to be affected by the random error that plagues single questions. An index, therefore, is a more reliable estimate of K than the results of a single item.[18]

---

[18] A quick way to compute an index is to add or average the results from several items or instruments. To produce a more credible and, therefore, useful instrument, it is a good idea to item analyze the different questions or instrument results which contribute to the index. The method for doing this is similar to that for constructing an attitude rating scale. Directions for computing indices and developing attitude rating scales can be found in Henerson, M. E., Morris, L. L., & Fitz-Gibbon, C. T. How to measure attitudes. In L. L. Morris (Ed.), *Program evaluation kit.* Beverly Hills: Sage Publications, 1978.

If a program plan, or perhaps a theory, has guided your examination of the program, then a particularly useful index for summarizing your findings at each site might be an estimate of *degree of implementation*. How you calculate such an index will vary with the setting. You would, however, select a set of the program's few most critical characteristics, and then compute the index from judgments of how closely the program depicted by the data from one or more instruments has put these into operation. The simplest index of degree of implementation would result from a checklist on which observers note *presence* or *absence* of important program components. The index would equal the number of *present* boxes checked.

Computing results for item by item interpretation. In addition to, or instead of, drawing relationships in your data, you may simply want to report results from your implementation instruments item by item. There are myriad ways to summarize and display this kind of data. Most of these are beyond the scope of this book, and you should consult a book on data analysis and reporting for more detailed suggestions.[19]

For the purpose of summarizing responses to individual items, you might want to present totals, percentages, or group averages. In some instances, computation will involve nothing more than adding tallies.

Example. Of the 50 children interviewed, 19 boys and 13 girls reported having taken part in the after-school recreation program. These 32 children reported having engaged in the following activities:

	boys	girls	total
handball	19	7	26
bars and rings	16	12	28
team games (baseball, kickball)	17	10	27
handicrafts	12	12	24
chess	8	8	16
checkers	10	8	18

---

[19] See in particular, Fitz-Gibbon, C. T., & Morris, L. L. How to calculate statistics; Morris, L. L., & Fitz-Gibbon, C. T. How to prepare an evaluation report; Henerson, M. E., Morris, L. L., & Fitz-Gibbon, C. T. How to measure attitudes. In L. L. Morris (Ed.), *Program evaluation kit*. Beverly Hills: Sage Publications, 1978.

In other cases, you may want to convert the numbers to percentages:

Example. Observers used a coded behavior record method to record teacher-student question-answer contacts during one week of lab periods in a high school chemistry program. From these extensive coded behavior records, the evaluator was able to find 503 teacher-student question-answer contacts. The evaluator classified these according to the following code:

t teacher
s student

q asks a question
r gives a response
o says nothing

According to this code, tq-sr-tq means that a teacher asked a question, a student gave a response, and the teacher asked another question. Accordingly, different sorts of conversation patterns, plus their relative frequencies, could be broken down as follows:

tq-sr-tq	*tq-sr-tr*	*tq-sr-to*	*sq-tq*	*sq-tr*	*sq-to*	*other*
90 (18%)	45 (9%)	42 (8%)	102 (20%)	80 (16%)	34 (7%)	110 (22%)

It was noted that the frequency of teacher questioning after student response was relatively high. This was a desirable behavior that the program had sought to foster.

If questions on the instrument demand answers that represent a *progression*, you may wish to report an *average answer* to the question.

```
What percent of the period did the teacher spend on disci-
pline?
☐ virtually none          ☐ about 75%
☐ about 25%               ☐ nearly all the time
☐ close to half
```

Averages can be graphed, displayed, and used in further data analyses. Be careful, however, to assure yourself that the average is truly representative of the responses that you received. If you notice that responses to a particular question pile up at two ends of the continuum, then the answers seem to be *polarized* and averages will not be representative. To report such a result by an average would be misleading to the audience.

3-24

Whether or not you become embroiled in reporting means and percentages ~~and looking for relationships~~, you will probably have to use the data you collect to underpin a *program description*. Program descriptions are usually presented as narrative accounts or descriptive tables such as Table 3, page 59 or Table 4, below.

TABLE 4
Project Monitoring--Activities[15]

Objective 6: By February 29, 19YY, each participating school will implement, evaluate results, and make revisions in a program for the establishment of a positive climate for learning.

Winona School District
Wiley School

Activities for this objective	19XX				19YY					
	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
6.1 Identify staff to participate		I	C							
6.2 Selected staff members review ideas, goals, and objectives		I	P	P	C					
6.3 Identify student needs		U	I	P	C					
6.4 Identify parent needs		U	I	P	C					
6.5 Identify staff needs		U	I	P	C					
6.6 Evaluate data collected in 6.3 - 6.5						I	U	C		
6.7 Identify and prioritize specific outcome goals and objectives			I	U	P	P	C			
6.8 Identify existing policies, procedures, and laws dealing with positive school climate		U	I	P	P	C				

Evaluator's Periodic Progress Rating:
I = Activity Initiated P = Satisfactory Progress
C = Activity Completed U = Unsatisfactory Progress

Table 4 best suits interim formative reports concerned with how faithfully the program's *actual schedule* of implementation conforms to what was originally planned. A formative evaluator can use this table to report, for instance, the results of monthly site visits to both the program director and the staff at each location. Each brief interim report consists of a table, plus accompanying comments explaining why ratings of "U," unsatisfactory implementation, have been assigned.

---

16. This table has been adapted from a formative monitoring procedure developed by Marvin C. Alkin.

TABLE 2 Methods for Collecting ~~Backup~~ Implementation Data

	Advantages	Disadvantages
Method 1: Examine Records. *Records* are systematic accounts of regular occurrences consisting of such things as attendance and enrollment reports, sign-in sheets, library checkout records, permission slips, counselor files, teacher logs, individual student assignment cards, etc.	• Records kept for purposes other than the program evaluation can be a source of data gathered without additional demands on people's time and energies. • Records are often viewed as objective and therefore credible. • Records set down events at the time of occurrence rather than in retrospect. This also increases credibility.	• Records may be incomplete. • The process of examining them and extracting relevant information can be time-consuming. • There may be ethical or legal constraints involved in your examination of certain kinds of records—counselor files, for example. • Asking people to keep records specifically for the program evaluation may be seen as burdensome.
Method 3: Conduct Observations. *Observations* require that one or more observers devote all their attention to the behavior of an individual or group within a natural setting and for a prescribed time period. In some cases, an observer may be given detailed guidelines about who or what to observe, when and how long to observe, and the method of recording the information. An instrument to record this kind of information would likely be formatted as a questionnaire or tally sheet. An observer may also be sent into a classroom with less restrictive instructions, i.e., without detailed guidelines, and simply asked to write a ~~[illegible]~~ naturalistic account of events ~~which~~ that occurred within the prescribed time period.	• Observations can be highly credible when seen as the report of what actually took place presented by disinterested outsider(s). • Observers provide a point of view different from that of people most closely connected with the program.	• The presence of observers may alter what takes place. • Time is needed to develop the observation instrument and train observers if the observation is highly prescribed. • It is necessary to locate credible observers if the observation is not carefully controlled. • Time is needed to conduct sufficient numbers of observations. • There are ~~usually~~ often scheduling problems.
Method 2: Use Self-Report Measures. *Questionnaires* are instruments that present information to a respondent in writing or through the use of pictures and then require a written response (a check, a circle, a word, a sentence, or several sentences).	• Questionnaires provide the answers to a variety of questions. • They can be answered anonymously. • They allow the respondent time to think before responding. • They can be given to many people, at distant sites, simultaneously. • They can be mailed. • They impose uniformity on the information obtained by asking all respondents the same things, e.g., asking teachers to supply the names of all math games used in class throughout the semester.	• They do not provide the flexibility of interviews. • People are often better able to express themselves orally than in writing. • Persuading people to complete and return questionnaires is sometimes difficult.
Interviews involve a face-to-face meeting or telephone discussion between two (or more) persons in which a respondent answers questions posed by an interviewer. The questions may be predetermined, but the interviewer is free to pursue interesting responses. The respondent's answers are usually recorded in some way by the interviewer during the interview, but a summary of the responses is generally completed afterwards.	• Interviews can be used to obtain information from people who cannot read and from non-native speakers who might have difficulties with the wording of written questions. • Interviews permit flexibility. They allow the interviewer to pursue unanticipated lines of inquiry.	• Interviewing is time-consuming, and can be hard to schedule. • Sometimes the interviewer can unduly influence the responses of the interviewee. • Interview data may be difficult to summarize.

TABLE 3

Program Ex-Cell Implementation Description

Program Component:
4th Grade Reading Comprehension--Remedial Activities

Person responsible for implementation	Target group	Activity	Materials	Organization for activity	Frequency/duration	Amount of progress expected
Teacher	Students	Vocabulary drill and games	SMA word cards, 3rd & 4th level	Small groups (based on CTBA vocabulary score)	Daily, 15-20 minutes	Completion of SMA, Level 4, by all students
			Teacher-developed word cards, vocabulary	Same	Same	None specified
			Old Maid	Same	Same	None specified
Teacher/Aide	Students	Language experience activities --keeping a diary, writing stories	Student notebooks, primary and elite typewriters	Individual	Productions checked weekly (Fridays); students work at self-selected times or at home	Completion of at least one 20-page notebook by each child; 80% of students judged by teacher or aide as "making progress"
Reading specialist/ teacher, student tutors	Students	Peer tutoring within class, in readers and workbooks	United States Book Company Urban Children reading series and workbooks	Student tutoring dyads	Monday through Thursday, 20-30 minutes	Completion of 1+ grade levels by 80% of students
Principal	Parents	Outreach--inform parents of progress; encourage at-home work in Urban Children texts; hold two Parents' nights; periodic conferences		All parents for program come to Parents' Night; other contact with parents on individual basis	Two Parents' Nights--Nov. and Mar.; 3 written progress reports in Dec., Apr., June; other contact with parents ad hoc	

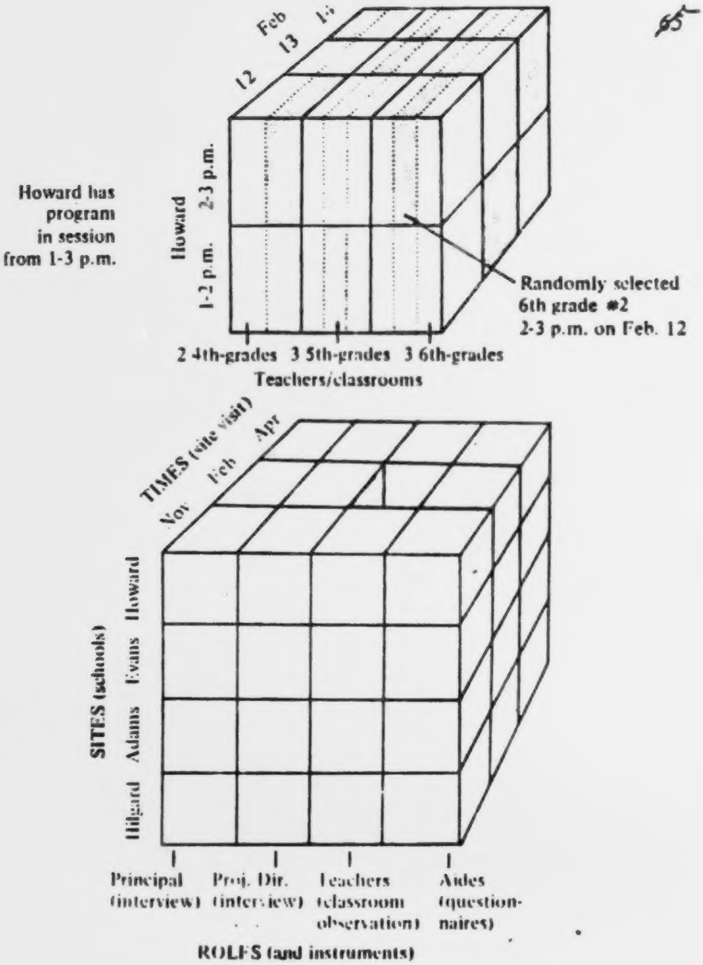

Feb
12
13
14
Howard has program in session from 1-3 p.m.
Howard
2-3 p.m.
1-2 p.m.
Randomly selected 6th grade #2 2-3 p.m. on Feb. 12
2 4th-grades
3 5th-grades
3 6th-grades
Teachers/classrooms
TIMES (site visit)
Nov
Feb
Apr
SITES (schools)
Howard
Evans
Adams
Hilgard
Principal (interview)
Proj. Dir. (interview)
Teachers (classroom observation)
Aides (questionnaires)
ROLES (and instruments)

Figure 2. Cubes depicting a sampling plan for measuring implementation of a middle-grades reading program in four schools within a district. The large 4x4x3 cube shows the overall data-collection plan from which sample cells may be drawn. The smaller cube shows selection of a random sample (shaded segments) of classrooms and reading periods chosen at Howard School for observation during a 3-day February site visit.

Example of a people-item data roster

Observation Response Form (results from classroom 1)

Implementation Objective: Students will direct and monitor their own progress in math activities. During the math period:	applies to most	applies to some	applies to [illegible]	applies to [illegible]
1. Students worked on individual math assignments.	4	3	2	1
2. Students asked for help with finding materials to work on.	1	2	3	4
3. Students loitered about, working at no activity in particular.	1	2	3	4
4. Students used self-testing sheets.	4	3	2	1
5. Students sought out aide [illegible] testing				

Summary Sheet (people-item format)

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	etc.
Classroom 1	3	4	4	3	4	4	
Classroom 2							
Classroom 3							

etc.

Examples of quick-tally sheets

Questionnaire

yes	no	uncertain	
☐	☐	☐	1. Were the materials available when you needed them?
☐	☐	☐	2. Were the materials suitable for your students

Summary Sheet (quick-tally format)

Item #	yes	no	uncertain
1	𝍸 ||||	|||	||
2	|	𝍸 𝍸	|

etc.

Observation Instrument

Rate the quality of the group interaction you observed using the scale provided.

1 - unsatisfactory
2 - poor
3 - so-so
4 - good
5 - outstanding

1 2 3 4 5 1. The harmony with which the working group functioned

1 2 3 4 5 2. The involvement of all members in contributing to group planning

Summary Sheet (quick-tally format)

Item #	1 unsatisfactory	2 poor	3 so-so	4 good	5 outstanding
1	|	|		|	𝍸 |
2		|	|||	𝍸	

etc.

Chapter 4

METHODS FOR MEASURING PROGRAM IMPLEMENTATION: PROGRAM RECORDS

An historian studying the activities of the past relies in large part on primary sources, documents created at the time in question that, taken together, allow the scholar to recreate a developmental picture of what happened. Evaluators, too, can take advantage of the historian's methods by using a program's records--the tangible remains of program occurrences--to construct a credible portrait of what has gone on in the program. Unobtrusive measures, methods of data-collection that, because they are ongoing or require little effort on any one person's part, can provide valuable information concerning program implementation. Consider the list of commonly kept records given in Table 5. Any of these could be used to develop your description of a program implementation, although you will most likely find the clearest overall picture of the program in those records that program staff have kept systematically on an ongoing basis.

If you want to measure program implementation by means of records, consider two things:

- How can you make good use of *existing* records?
- Can you *set up* a record-keeping system that will give you needed information without burdening the staff?

Where records are already being kept, you can use them as a source of information about the activities they are intended to record. Since the progress charts, attendance records, enrollment forms, and the like kept for the program will seldom cover all you need to know, ~~though~~, you might try to arrange for the staff or students to maintain additional ~~documents~~. Of course, you will be able to *set up* record-keeping only if your evaluation begins early enough during program implementation to allow for an accurate picture of what has occurred.

In most cases, it is ~~not~~ realistic to expect that the staff will keep records over the course of the program solely to help *you* gather implementation information, unless these records are easy to maintain (e.g., parent-aide sign-in sheets) or are useful for their own purposes as well. You will do best if you come up with a valid reason why the staff should keep records, and attempt to align your information needs with theirs. You could, for instance, gain access to records by offering a service.

4-2

Table 5. Records Often Produced by Educational Programs

- Certificates upon completion of activities
- Completed student workbooks
- Student assignment sheets
- Dog-eared and worn textbooks
- Products produced by students (e.g., drawings, lab reports, poems, essays)
- Attendance and enrollment logs
- Sign-in and sign-out sheets
- Progress charts and checklists
- Unit or end-of-chapter tests
- Teacher-made tests
- Circulation files kept on books and other materials
- Diplomas and transcripts
- Report cards
- Letters of recommendation

- Activity or field-trip rosters
- Letters to and from parents, business persons, the community
- Letters of recommendation
- Logs, journals, and diaries kept by students, teachers, or aides
- Parental permission slips
- In-house memos
- Flyers announcing meetings
- Records of bookstore or cafeteria purchases or sales
- Legal documents (e.g., licenses, insurance policies, rental agreements, leases)
- Bills, purchasing orders, and invoices from commercial firms providing goods and services
- Minutes or tape-recordings of meetings
- Newspaper articles, news releases, and photographs
- Standardized test scores (local and state)

Table 5.

4-3

For example, by agreeing to write software for a custom-made management information system, the evaluator of an adolescent parenting center structured ongoing data collection of value ~~both~~ to [illegible] his clients, and to any future evaluator. In another instance, an evaluator was able to monitor program implementation at school sites statewide by helping schools write the periodic reports that had to be submitted to the State Department of Education.

Implementation Evaluation Based On Already Existing Records [illegible]

The following is a suggested procedure to help you find and extract pertinent information within the ~~program's already existing~~ records, and ~~to extract that information~~. that already exist in a program.

Step 1. Construct a program characteristics list

Compose a list of the materials, activities, ~~and~~/or administrative procedures about which you need ~~backup~~ supporting data. This procedure was detailed in Chapter 2, pages [illegible] to [illegible].

Step 2. Find out from the staff or the program director what records have been kept and which of these are available for your inspection.

Be sure you are given a complete listing of every record that the program produced, whether or not it was kept at every site. Probe and suggest sources that might have been forgotten. Draw up a list of all records that will be available to you.

[illegible]

If part of your task is to show that the program as implemented represents a departure from past or common practice, you might include records kept *before* the program.

Step 3. Match the lists from Steps 1 and 2

For each type of record, try to find a program feature about which the record might give information. Think about whether any particular record might yield evidence of the following:

- The *duration* or *frequency* of a program activity
- The *form* that the activity took, what it typically looked like; you will find this information only in narrative records such as curriculum manuals and logs, journals, or diaries kept by the participants
- The extent of student or other participant involvement in the activities—attendance, ~~good behavior, etc.~~ conduct, interest, etc.

Do not be surprised if you find that few available records will give you the information you need. The program staff has maintained records to fit its own needs; only *sometimes* will these overlap with yours.

BEST COPY

Step 4. Prepare a sampling plan for collecting records

General principles for setting up a data collection sampling plan were discussed in Chapter 3, page 60. The methods described there direct you either to sample typical *periods* of program operation at diverse sites, or to look intensively at randomly chosen cases. Were you to use the former method for describing, say, a language arts program, you might ask to see "library sign-in sheets and circulation files for the *fall quarter* at ~~*Baxter*~~ Boynton *Junior High*," as well as for other times and other places, all randomly chosen. The latter method directs that you focus on a few sites in detail. An intensive study might cause you to choose ~~Baxter~~ Boynton as representative of participating junior high schools and examine its whole program *in addition to* the library component. You could, as well, find your own way to mix the methods.

If part of your program description task involves showing the extent to which the program is a departure from usual practice, you could include in the sample *sites not receiving the program* and use these for comparison.

Step 5. Set up a data collection roster, and plan how you will transfer the data from the records you examine

The data roster for examining records should look like a questionnaire –

"How many people used the library during this particular time unit?"

"How long did they stay?" "What kinds of books did they check out?"

Responses~~, entered by your data transfers,~~ can take the form of tallies or answers to multiple choice questions.

When data collection is complete, you might still have to transfer it from the multitude of rosters or questionnaires used in the field, to single data summary sheets, described in Chapter 3, pages 67 to 71.

Step 6. Where you have been able to identify available records pertinent to examining certain program activities, set up a means for obtaining access to those records in such a way that you do not inconvenience the program staff.

Arrange to pick up the records or copy them, extract the data you need, and return them as quickly and with as little fuss as possible. A member of the evaluation staff should fill out the data summary sheet; *program* staff should not be asked to transfer data from records to roster.

Setting Up a Record-Keeping System

What follows is a suggested procedure for establishing a record-keeping system or what is sometimes called a management information system (MIS). With the growing availability of computers for even small organizations, program personnel increasingly have the capacity of collect and maintain data for use on an ongoing basis. While evaluators seldom have the luxury of building provisions for their own record-keeping into the program itself, they should be prepared to take advantage of ~~the~~ such an opportunity if it arises, remembering to structure the record-keeping primarily to the needs of program staff and planners and only secondarily to the needs of the implementation evaluation.

Step 1. Construct a program characteristics list for each program you describe.

Compose a list of the materials, activities, ~~and/~~or administrative procedures about which you need supporting data. (This procedure was outlined on pages ?? to ??.) If the evaluation uses a control group design or if one of your tasks is to show that the program represents a departure from usual practice in the district, you may need to describe the implementation of more than one program. You should construct a separate list of characteristics for each program you describe.

Attach to your list, if possible, columns headed in the manner of Columns 2 and 3, Table 6, page 83. This table has been constructed to accompany an example illustrating the procedure for setting up a record-keeping system.

Step 2. Find out from the program staff and planners which records will be kept during the program as it is currently planned

Be sure this list includes tests to be given to students, reports to parents, assignment cards--all records that will be produced over the course of the program. Check the list of items in Table 5 for their availability.

Step 3. For each program characteristic listed in Step 1, decide if a proposed record can provide information that will be both useful and sufficient for the evaluation's purposes

First, examine the list of records that will be available to you. Will any of them be useful as a check of either quantity, quality, regularity of occurrence, frequency, or duration of the program characteristic? If it will, enter its name on your activities chart next to the activity whose occurrence it will demonstrate. Jot down a judgment of whether the record *as is* will fit your needs or whether it might need slight modification. Also enter the number of collections or updatings of the record that will take place over the course of the program. If the number of collections seems insufficient to give a good picture of the program, talk to the staff to request more fr[illegible]t updating.

4-6

Step 4. For those characteristics that are not covered by the staff's list of planned records, decide if simple additions or alterations can provide appropriate and adequate evaluation data

When you have finished your review of records that will be available, look closely at the set of program activities about which you still need information. These will *not* be covered by the staff's list of planned records. Try to think of ways in which alteration or simple addition to one of the records already scheduled for collection might give you information on the frequency of occurrence or form of one of the activities on your list. If it appears that slight alteration of a record will give you the information you need, note the name of the record and its planned collection frequency and request that the program staff make the change you need.

Step 5. Meet with program staff first to review the planned records that will provide data for the evaluation and second to recommend changes and additions for their consideration

Before seriously approaching the staff and asking for their assistance with your information collection plan, however, scrutinize it as follows:

- Will it be too time-consuming for the staff to fill out regularly?
- Will the staff members perceive it as useful to *them*?
- Can you arrange a feedback system of any sort to give the staff *useful* information based on the records you plan to ask them to keep?

If the information plan you have conceived passes these checkpoints, suggest it to the staff.

Try to avoid data overload. Do not produce a mass of data for which there is little use. The way to avoid collecting an unnecessary volume of data is to plan data use before data collection.

BEST COPY

4-7

Step ~~4~~6. Prepare a sampling plan for collecting records

Once you know which records will be kept to facilitate your implementation evaluation, decide where, when, and from whom you will collect them. General principles for setting up a data collection sampling plan were discussed in Chapter 3, page 60. The methods described there produce two types of samples:

- A sample that selects typical time periods or episodes from the program at diverse sites
- A sample that selects people, classes, schools or other sites, considering each *case* typical of the program

Your sampling plan could use either or both.

Step ~~5~~7. Set up a data collection roster and plan how you will transfer data from the records you examine

The data roster for examining records should resemble a questionnaire for which answers take the form of tallies or, in some cases, multiple-choice items.

The data roster is a means for making implementation information accessible to you when you need it so that it can be included in the data analysis for your report. The roster, you will notice, compiles information from a *single* source, covering a single time period. For the purpose of your report, you will usually have to transfer all of the roster data to a *data summary sheet* in order to look at *the program as a whole.* Chapter 3 describes data summary sheets, including those for managing data processing by computer, beginning on page 67.

Step ~~6~~8. Set up a means for obtaining easy access to the records you need

Gather records from the staff in a way that minimally interferes with their busy work schedules. You, or your delegate, should arrange to collect workbooks, reports, checklists, or whatever, photocopy them or extract the important data, and return these records as quickly as possible. Only in those rare situations where the staff itself is ungrudgingly willing to participate in your data collection should you ask them to bring records to you or transfer information to the roster.

Step 9. Check periodically to make sure that the information you have requested from program staff is in fact being recorded accurately and completely

It is one thing for an evaluator to plan an implementation evaluation thoroughly at the beginning of a program. It is another thing altogether ~~for an evaluator~~ to return, say, a year later and actually find the records ready for her use. In many cases you may return at the end of the year to discover that what you thought program staff were going to do in the area of record keeping and what they actually did were two different things. If the effectiveness of your evaluation relies on records kept by program personnel, you are well advised to check periodically to make sure that the information you need is being collected and maintained. In the press of program activities, record-keeping may become burdensome or, given limited resources, even an inappropriate use of staff time.

Chpt 4 Example
(no changes)

4-9

Example. Ms. Gregory, Director of Evaluation for a mid-sized school district, is intending to evaluate the implementation of a ~~s~~tate-funded compensatory education program for grades K through 3. The program uses individualized instruction. After examining the program proposal and discussing the program with various staff members, she has constructed the implementation record-keeping chart shown in Table 6.

TABLE 6
Example of an Implementation Record-Keeping Chart

Column 1 Activities (3rd grade)	Column 2 Record to be used for monitoring the activity--adequate for assessing implementation?	Column 3 Frequency and regularity of record collection--sufficiently representative to assess implementation?
1) Early morning warm-up, group exercise (10 min./day)		
2) Individualized reading (45 min./day)		
Each student:		
a) reading aloud with teacher/aide (3 times/week)		
or		
b) reading cassette work at recorder center (3 times/week)		
or		
c) reading seatwork--choice of workbook or library book		
3) Perceptual-motor time (15 min./day in school gym with aide)		
Two parts:		
a) clapping rhythm exercise (in group)		
b) open balance period (individual, on jungle gym, balance beam, etc.)		

Example continued. Ms. Gregory found that program teachers *already planned* to keep records of students' progress in "reading aloud" (Activity 2a) and of their work with audio tapes in the "recorder corner" (2b). Further, this record collection *as planned* seemed to Ms. Gregory to give her exactly the implementation information she needed: teachers planned to monitor reading via a checksheet that would let them note the date of each student's reading session and the number of pages read.

Teachers also planned to note the quality of student performance, a bit of data that Ms. Gregory did *not* need. Work with cassettes in the recorder corner (2b) was to be noted on a special form by an aide, but *only* the progress of children with educational handicaps would be recorded. These audio corner records, Ms. Gregory decided, would not be adequate. She needed data on *all* children's use of the tapes. She noted the usefulness of this information on her chart, with an additional notation to speak to the staff about changing record-keeping in the recorder center to include at least a periodic random sample from the whole class.

Column 1 Activities	Column 2 Record	Column 3 Collection
a) reading aloud with teacher/aide (3 times/week)	teacher/aide's record book: gives dates of recording, no. of pages read--adequate	constant recording--adequate
or		
b) reading cassette work at recorder center (3 times/week)	aide's recording form; gives amount of time, progress, distractions--adequate	only on EH children--inadequate: speak with staff; could they look at all students?

BEST COPY

4-10

Example continued. Ms. Gregory needed some information for which no records were planned. For instance, teachers and aides did not intend to keep records of students' participation in "perceptual-motor time" (Activity 3). Ms. Gregory noted this and determined to meet with the staff to suggest some data collection.

Column 1 Activities	Column 2 Record	Column 3 Collection
3) Perceptual-motor time (15 min./day in school gym with aide)	none--inadequate: suggest that aide keep a checklist or diary of length and content of daily sessions	
Two parts:		
a) clapping rhythm exercise (in group)	none--inadequate: aide diary?	
b) open balance period (individual, on jungle gym, balance beam, etc.)	none--inadequate: aide diary?	

Ms. Gregory spoke with aides about the possibility of keeping a diary of perceptual-motor activities. Aides resisted this idea; they wanted the period to be relatively undirected, and they saw it as a break for themselves from regular in-class record-keeping. They did, however, feel that it would be useful to them to have a record of each student's progress in balancing and climbing. Ms. Gregory was thus able to persuade them to construct a checklist called GYM APPARATUS I CAN USE, to be kept by the students themselves and collected once a month. Ms. Gregory decided to collect data on the "clapping" part of the perceptual-motor period in some way other than by examining records, perhaps via a questionnaire to aides at the end of the year, or through observations.

Example continued. Ms. Gregory was faced with the responsibility of practically single-handedly evaluating a comprehensive year-long program. As it turned out, Ms. Gregory was quite successful at finding records that would provide her with the implementation information she needed. The following records would be made available to her:

- The teachers' record books showing progress in read-aloud sessions
- Aides' recording forms of students' recorder corner work
- Students' GYM APPARATUS I CAN USE checklists

Also available were other records for teaching math, music, and basic science—topic areas not included in the example. All records would be available to Ms. Gregory throughout the year. But how would she find time to extract data from them all?

By means of a time sampling plan, Ms. Gregory could schedule her record collection and data transcription to make the task manageable. First, she chose a *time unit* appropriate for analyzing the types of records she would use. The teachers' records of read-aloud sessions, for example, should be analyzed in *weekly* units rather than daily units. According to Ms. Gregory's activities list, the program did not require students to read every day; they *must* read for the teacher at least three times *per week*. Perceptual-motor time could be analyzed by the day, however, since the program proposal specified a *daily* regimen. She then selected a random sample of *weeks* from the time span of the program and arranged to examine program records at the various sites. She selected *days* for which gym apparatus progress sheets would be examined.

Site and participant selection was random throughout. For each week of data collection, she randomly chose four of the eight participating schools, and within them, two classes per grade whose records would be examined.

Example continued. Having sampled both time units and classrooms, Ms. Gregory consulted teachers' records from eight classrooms at each grade level for the week of January 26. Once she had prepared a list of the 30 students in one of the third grade samples, she

- Tallied the number of times each one read
- Recorded the number of pages read
- Calculated the mean number of pages read that week per student

Ms. Gregory's data roster for gathering information on third-grade read-aloud sessions from one teacher's record book looked like Table 7.

TABLE 7
Example of a Data Roster
for Transferring Information
From Program Records

Individualized Program

Class: Mr. Roberts--3rd Grade
School: Allison Park
Activity: Reading aloud with teacher or aide
Data source: Teacher's record book
Questions: How often did children read per week? How many pages did they cover?
Time Unit: Week of January 26

Student	Tally of times student read	No. of pages read	Mean no. of pages read
Adams, Oliver	//// 4	4, 5, 6, 5	5
Ault, Molly	// 2	3, 4	3.5
Caldwell, Maude	/// 3	4, 3, 5	4
Connors, Stephen	~~////~~ 5	1, 4, 6, 5, 4	4
Ewell, Leo	/// 3	3, 5, 4	4
Goldwell, Nora	~~////~~ 5	6, 2, 3, 4, 5	4
Gross, Joyce	// 2	7, 8	7.5

Chapter 5

METHODS FOR MEASURING PROGRAM IMPLEMENTATION: SELF-REPORTS

Chapter 4 described ways in which evaluators can use program records to provide one type of implementation information. Because records are for the most part written documents, however, the picture they ~~help~~ create may be incomplete, lacking the details that only those who experienced the program can provide. A good way to find out what a program actually looked like is to ask the people involved, ~~and~~ The focus of this chapter, therefore, is self-reports, the personal responses of program faculty, staff, administration, and participants.

Self-reports typically take one of two forms: questionnaires and interviews. Questionnaires asking about different individuals' experiences with a program enable one evaluator to collect information efficiently from a large number of people. Individual or group interviews are more time-consuming, but provide face-to-face descriptions and discussion of program experiences.

Where there is a plan or theory ~~describing~~ for the program, gathering information from staff will involve questioning them about the consistency between program activities as they were planned and as they actually occurred. Where the program has not been prescribed, information from people connected with it will [illegible] describe how the program evolved.

Whether they are questionnaires or interviews, self-reports also differ on the dimension of time. They can consist either of periodic reports throughout the program or retrospective reports after the program has ended.

(no 4) Periodic reports will generally yield more accurate implementation information because they allow respondents to report about program activities soon after they have occurred, when they are still fresh in memory. For this reason, they are nearly always more credible than retrospective reports. Periodic reports should be used even when your role is summative and you are required to describe the program only once, at its conclusion.

Retrospective self-reports should be used in only two cases: when there is no other choice (e.g., because the evaluation is commissioned near the program's conclusion) or when the program is small enough or of such short duration that reconstructions after-the-fact will be believable. What follows are step-by-step directions for collecting self-reports through periodic questionnaires or interviews. These can be adapted easily to develop ~~create~~ a retrospective report.

How To Gather Periodic Self-Reports Over the Course of the Program

sp'

Step 1. Decide how many times you will distribute questionnaires or conduct interviews and from whom you will collect these self-reports

As soon as you begin working on the evaluation and as early as possible in the program's life, decide how often you will need to collect self-report information. This decision will be determined by three factors:

- The homogeneity of program activities. If each program unit has essentially the same format as the others, then you will not need to document descriptions of particular ones. If, for example, a company's program for updating employees' knowledge in a technical field consists of standardized lessons containing a lecture, reading, and class discussion, then any one lesson you ask about at any given site will reflect the typical format of the program. In such a case you can plan data collection at your discretion. If, on the other hand, the program has certain unique features, say group project assignments that will vary from site to site or special guest lectures by local university professors, you will want to ask about these distinguishing program features as soon as they occur. This will give you a chance to digest information and provide immediate formative feedback to program planners and staff.

- *Your assessment of people's tolerance for interruptions.* Unless the program is sparsely staffed, you should not ask for more than three reports from any one individual over the span of a long-term program (e.g., a year). You can *sample*, of course, so that the chances are reduced that any one person will be asked to report often.
- *The amount of time you expect to have available for scoring and interpreting information in reports.*

Once you have decided when to collect self-reports, create a sampling strategy (see pages 60 to 64) by deciding whom you will ask for self-report information (both by title and by name) and how you will insure that various program sites are adequately represented.

Step 2. ~~Warn~~ Alert people that you will be requesting periodic information

As early during the evaluation as possible, inform staff members and others that in order to measure implementation of their program, you must ask that they provide you with information about how the program looks in operation.

Step 3. Construct a program characteristics list

Procedures for listing the characteristics of the program—materials, activities, administrative arrangements—that you will examine are discussed in Chapter 3, pages 57 to 60

Step 4. Decide—if you have not already—whether to distribute questionnaires, to interview, or to do both

You probably know about the relative advantages and disadvantages of using questionnaires or interviews. Table C, page 54 reminds you of some of them. If you are using self-report instruments to supplement program description data from a more credible source—observations or records—then questionnaire data should be sufficient. On the other hand, if self-report measures will provide your only implementation backup data, then you should interview some participants. ~~Unless~~ Even if you are a clever questionnaire writer, you probably cannot find out all you need to know about the program from a pencil-and-paper instrument, and interviews allow a sensitive evaluator to come face to face with important program concepts and issues.

Step 5. Write questions based on the list from Step 3 that will prompt people to tell you what they saw and did as they participated in the program

Anyone who writes questions or develops items on a regular basis would do well to consult the books listed at the end of the chapter as what can be presented here represents only a small part of available knowledge on how to do this well. The development of good items for questionnaires and interviews clearly combines art, science, common sense, and practice. What follows is a brief summary of things to consider when writing questionnaire or interview items. You should also review Table 8 for a list of pointers to follow when writing questions for a program implementation instrument.

To begin, one thing you will need to know is how participants used the materials and engaged in the activities that comprised the program. To this end, you should ask about three topics:

a. The *occurrence, frequency*, or *duration* of activities. Whether you collect frequency and duration information in addition to occurrence will depend on the program. To describe a Science Lab program, for instance, you would need merely to determine whether the planned labs occurred at all–and in the correct sequence. If, on the other hand, the program in question consisted of daily, 45-minute English conversation drills, then you would need to know whether the activity occurred with the prescribed frequency and duration.

b. The *form* the activities took. Gathering information on the form of the activities means asking about which students took part in the activities, which materials were used and how often, what activities looked like, and possibly where they occurred. It will also be useful to check whether the form of the activities remained constant or whether the activities changed from time to time or from student to student.

c. The *amount of involvement* of participants in these activities. Besides knowing what activities occurred, you should make some check on the extent of interest and participation on the part of the target group–say, the students. Even if activities were set up using the prescribed schedule, students can only be expected to have learned from them if they engaged the students' attention. Were students in a math tutoring program, for instance, mostly working on the prescribed exercises, or were they conversing about sports and clothes some of the time? Were students in an *unstructured* period actually exploring the enrichment materials, or were they just doing their homework? Some of this slippage is inevitable in every program (as in all human endeavor). Still, it is important to find out the extent of non-involvement in the program you are evaluating.

If you ~~*intend to*~~ are *design*ing *a questionnaire*, then you have a choice of two question formats: a closed (*selected*) or open (*constructed*) response format. Ease of scoring and clear reporting lead most evaluators to use *closed-response* questionnaires. On such a questionnaire, the respondent is asked to check or otherwise indicate a *pre-provided answer* to a specific question. Recording the answers involves a simple tally of response categories chosen. On the *open-response* questionnaire, the respondent is asked to write out a short answer to a more general question. The open-response format has the advantage of allowing respondents to freely give information you had not anticipated, but it is time-consuming to score; and unless you have available a large number of readers, it is not practical for any but the smallest evaluations. Most questionnaires ask principally closed-response questions, but add a few open-response options. These allow respondents to volunteer information important to the evaluation but not specifically requested.

To demonstrate how different question types result in different information, Figures 13, 14, and 15 present combinations of open- and closed-~~ended~~ response questions for collecting implementation information on the same program. Figure 13 is entirely open-ended; Figure 14 combines open- and closed-ended questions; and Figure 15 uses a closed-response format exclusively. While the data that would result from the questionnaire in Figure 15 would be easily analyzed, this ease is gained at the expense of the more detailed information that individual teachers ~~could~~ might write in on the two other questionnaire formats. The appropriateness of the questionnaire items finally selected will depend both on the questions asked in the evaluation as a whole and on the availability of evaluation staff to analyze open-response format items. In general, it is worth including at least one open-ended question on every questionnaire, whether or not the results will later be reported. Giving people an opportunity to write down their concerns alerts them to the importance of their perspective and provides the evaluation helpful information for guiding ~~later~~ program activities.

Like questionnaires, interviews can also take several forms, again depending on how questions are asked. Interviews can range from informal personal conversations with program personnel at one extreme to highly quantitative interviews that consist of a respondent and an evaluator completing a closed-response format questionnaire together at the other extreme. (Because this quantitative interview format doesn't take advantage of the face-to-face interaction of evaluator with respondent, it is more properly considered the enactment of a questionnaire, than an interview.)

~~For most interviews,~~ A basic distinction can be made between ~~those~~ [illegible] that are structured and those that are unstructured. In a structured interview, an evaluator asks specific questions in a pre-specified order. Neither the questions nor their order is varied across interviewers, and in its purest form the interviewer's job is merely to ask the predetermined questions and to record the responses. In cases where an evaluator already has ideas about how ~~the~~ a program looked, structured interviews can readily provide corroboration and supporting data.

By contrast, an unstructured interview can explore areas of implementation that were unplanned or that evolved differently from the plan. In an unstructured interview the evaluator poses a few general questions and then encourages

~~the~~ respondents to amplify ~~his~~ their answers. The unstructured interview is more like a conversation and does not necessarily follow a specific question sequence.

Unstructured interviews require considerable interviewing skill. General questions for the unstructured interview can be phrased in several ways. Consider the following questions:

- *How often, how many times, or hours a week did the program (or its major features) occur?*
- *What can you tell me about how the activities actually looked--can you recall an instance and describe to me exactly what went on?*
- *How involved did the students seem to be—did all students participate, or were there some students who were always absent or distracted?*
- *I understand that you are attempting to implement a behavior modification, or open classroom, or values clarification program here. What kinds of classroom activities have been suggested to you by this point of view?*

Since unstructured interviews resemble conversations and can easily go off track, they require not only that you compose a few questions to stimulate talk, but also that you write and use probes. Probes are short comments to stimulate the respondent to say or remember more and to guide the interview toward relevant topics. Two frequently used probes are the following:

Can you tell me more about that?

Why do you think that happened?

There is no set format for probes. In fact, a good way of probing to gain more complete information from respondents who have forgotten or left something out of their answer might be a simple:

I see. Is there anything else?

You should insert probes whenever the respondent makes a strong statement in either an expected or an unexpected direction. For instance, a teacher might say:

Oh, yes. Participation, student involvement was very high--100%.

The best probe for such a strong response is a simple rephrasing and repetition:

Your statement is that every student participated 100% of the time?

This probe leads the respondent to reconsider.

Step 6. Assemble the questionnaire or interview instrument

Arrange questions in a logical order. Do not ask questions that jump from one subject to another.

Compose an introduction. The introduction honors the respondents' right to know why they are being questioned. *Questionnaire* instructions should be specific and unambiguous. Aim for as simple a format as possible. You should assume that a portion of the respondents will ignore instructions altogether. If you feel the format might be confusing, include a conspicuous sample item at the beginning. Instructions for a mailed questionnaire should mention a *deadline* for its return, and you should enclose a self-addressed, stamped envelope.

Instructions for an interview can be more detailed, of course, and should include reassurances to ~~dilute~~ allay the respondent's initial apprehension about being questioned. Specifically, the interviewer should:

- *State the purpose of the interview.* Explain what organization you represent, and why you are conducting the evaluation. Explain the purpose of the interview. Describe the report you will have to make regarding the activities that occurred in the program: explain if possible how the information the respondent gives you might affect the program.
- State ~~whether~~ or not ~~If~~ *the respondent's statements can be kept confidential*, ~~[illegible]~~. In situations where a social or professional threat to the respondent may be involved, confidentiality of interviews must be stressed and maintained.
- *Explain to the respondent what will be expected during the interview.* For instance, if it will be necessary for the respondent to go back to the classroom to get records, explain the necessity of this action.

Some of the above information should probably be made available to questionnaire respondents as well. This can be done by including a cover letter with the questionnaire.

Step 7. Try out the instrument

Before administering or distributing any instrument, check it out. Give it to one or two people to read aloud, and observe their responses. Have the people explain to you their understanding of what each question is asking. If the questions are not interpreted as you intended, alter them accordingly.

Always rehearse the interviews. Whether you choose to prepare a structured or unstructured interview, once the questions for the interview are selected, the interview should be rehearsed. You, and other interviewers, should run through it once or twice with whoever is available a ~~wife~~ spouse,

~~a husband~~, an older child, a ~~secretary~~ colleague. This dry-run is a test of both the instrument and the interviewer. Look for inconsistency in the logic of the question sequencing and difficult or threateningly worded questions. Advise the person who is playing the role of respondent to be as uncooperative as possible to prepare interviewers for unanticipated answers and even hostility.

Step 8. Administer the instrument according to the sampling plan from Step 1

If you *mail questionnaires*, give respondents about two weeks to return them. Then follow up with a reminder, a second mailing, or a phone call if possible. How do you do such a follow-up if people are to respond anonymously? One procedure is to number the return envelopes, check them off a master list as they are returned, remove the questionnaires from the envelopes, and throw the envelopes away.

When distributing any instrument, ask administrators to lend their support. If the instrument carries the sanction of the project director or the school principal, it is more likely to receive the attention of those involved. The superintendent's request for quick returns will carry more authority than yours.

If you *interview*, consider the following suggestions:

- Interviewers should be aware of their influence over what respondents say. Questions about the administration of the program may be answered defensively if staff members fear their answers might make them look bad in a report. Explain to the respondents that the report will refer to no one personally. Understand, as well, that respondents will speak more candidly to interviewers whom they perceive as being like themselves—not representatives of authority.
- Interviewers should have a plan for dealing with reluctant respondents. The best way to overcome resistance is to be explicit about the interview and what it will demand of the respondent.
- If possible, interviews, particularly unstructured ones, should be recorded on audiotape to be transcribed at a later time. Recorded interviews enable you to summarize the information using exact quotations from the respondent; they also require a lot of transcription time. Transcribing the tape in full will take ~~at least half again as long~~ two or three times as long as the interview itself. An alternative is that interviewers *take notes* during an unstructured interview. Notes should include a general summary of each response, with key phrases recorded verbatim. If possible, summaries of unstructured interviews should be returned to respondents so that misunderstandings in the transcription can be corrected.

Step 9. Record data from questionnaires and interview instruments on a data summary sheet

Chapter 3, page 67, described the use of a data summary sheet for recording data from many forms in one place in preparation for data summary and analysis. Data from closed-response items on questionnaires and structured interview ~~schedules~~ can be transferred directly to the data summary sheet. Responses to open-response items and unstructured interviews will have to be summarized before they can be further interpreted. Procedures for reducing a large amount of narrative information by either summarizing or quantifying it were discussed on pages 71 to 73. Even if you plan to write a narrative report of your results, the data summary sheet will show trends in the data that can be described in the narrative.

BEST COPY

TABLE 8

Some Principles To Follow When Writing Questions For An Instrument To Describe Program Implementation

---

To ensure usable responses to implementation questions:

1. When possible, ask about specific—and recent—events or time periods such as *today's math lesson, Thursday's field trip, last week.* This persuades people to think concretely about information that should still be fresh in memory. To alleviate your own and the respondent's concern about representativeness of the event, ask for an estimate, and perhaps an explanation, of its typicality.

2. When asking a closed-response question, try to imagine what could have gone wrong with the activities that were planned. or ask program staff for their help. Use these possibilities as response alternatives. Resourceful anticipation of likely activity changes will affect the usefulness of the instrument for uncovering changes that did indeed occur. If you feel that you cannot adequately anticipate discrepancies between planned and actual activities, then add "other" as a response alternative and ask respondents to explain.

3. Be sure that you do not *answer* the question by the way you ask it. A good question about what people *did* should not contain a suggestion about how to answer. For instance, questions such as "Were there ~~4th and 6th graders~~ new hires in the program?" or "Did you meet every Monday afternoon?" suggest information you should receive from the respondent. Rather, these questions should be phrased, "What were the ~~grade~~ experience levels of the ~~students~~ participants in the program?" "What days of the week and how regularly did you meet?"

4. Identify the frame of reference of the respondents. In an interview, you can learn a great deal from how a person responds as well as from what he says; but when you use a questionnaire, your information will be limited to written responses. The *phrasing* of the questions will therefore be critical. Ask yourself:

   - *What vocabulary would be appropriate to use with this group?*
   - *How well informed are the respondents likely to be?* Sometimes people are perfectly willing to respond to a questionnaire, even when they know little about the subject. They feel they are *supposed* to know, otherwise you would not be asking them. To allow people to express ignorance gracefully, you might include lack of knowledge as a response alternative. Word the alternative so that it does not demean the respondent, for instance, "I have not given much thought to this matter."
   - *Does the group have a particular perspective that must be taken into account—a particular bias?* Try to see the issue through the eyes of the respondents before you begin to ask the questions.

BEST COPY

(add [illegible])

DOCUMENTATION QUESTIONNAIRE

Peer-Tutoring Program

The following are questions are the peer-tutoring program implemented this year. We are interested in knowing your opinions about what the program looked like in operation. Please respond to each question, and feel free to write additional information on the back of this questionnaire.

1. How was the peer-tutoring program structured in your classroom?

2. How were tutors selected?

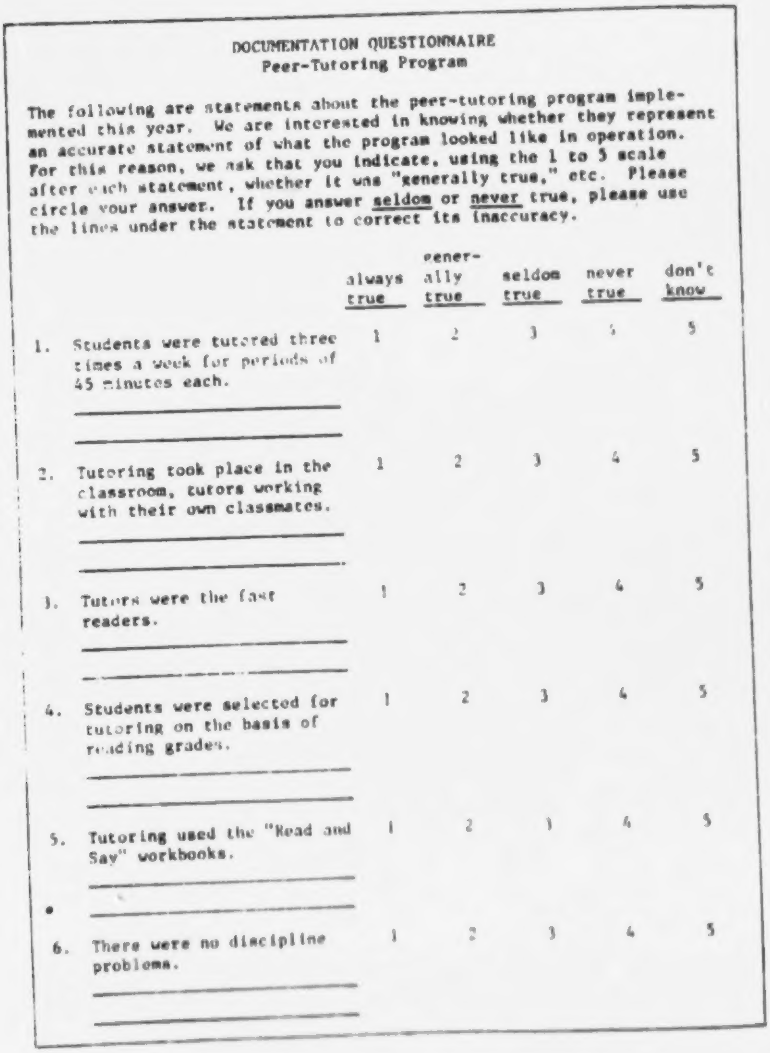
3. How were students selected for tutoring?

4. What materials seemed to work best in the peer-tutoring sessions? Why?

5. What were the strengths of the peer-tutoring program this year?

6. What changes would you make to improve the program next year?

Figure 13. Example of an open response questionnaire.

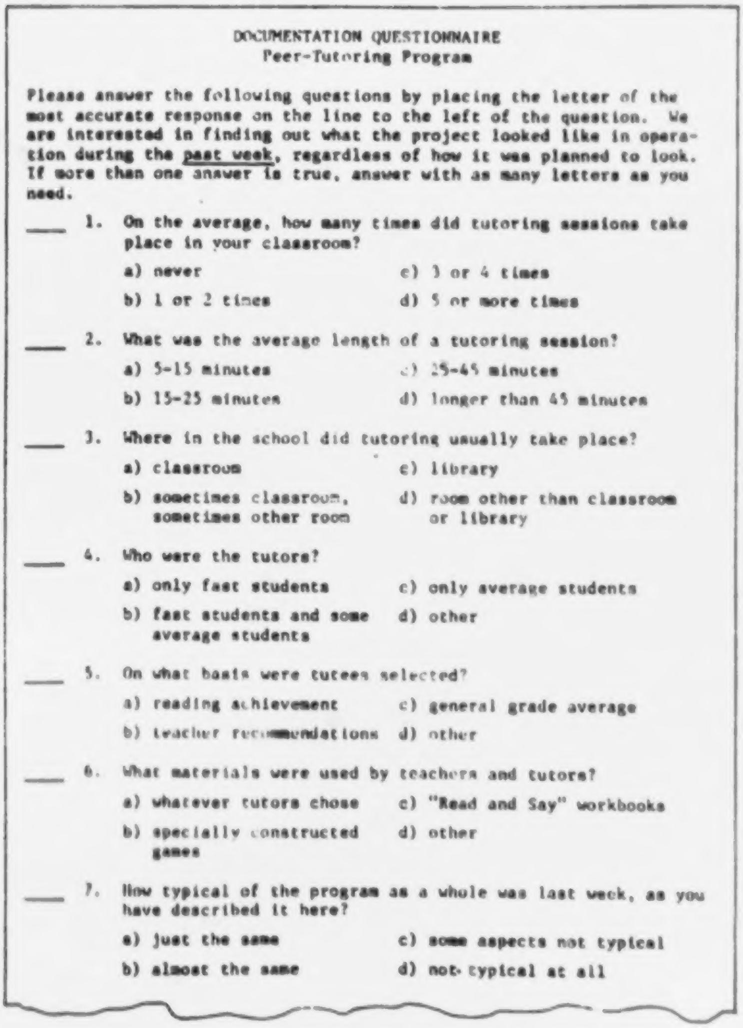


DOCUMENTATION QUESTIONNAIRE
Peer-Tutoring Program

The following are statements about the peer-tutoring program implemented this year. We are interested in knowing whether they represent an accurate statement of what the program looked like in operation. For this reason, we ask that you indicate, using the 1 to 5 scale after each statement, whether it was "generally true," etc. Please circle your answer. If you answer seldom or never true, please use the lines under the statement to correct its inaccuracy.

	always true	generally true	seldom true	never true	don't know
1. Students were tutored three times a week for periods of 45 minutes each.	1	2	3	4	5
2. Tutoring took place in the classroom, tutors working with their own classmates.	1	2	3	4	5
3. Tutors were the fast readers.	1	2	3	4	5
4. Students were selected for tutoring on the basis of reading grades.	1	2	3	4	5
5. Tutoring used the "Read and Say" workbooks.	1	2	3	4	5
6. There were no discipline problems.	1	2	3	4	5

Figure 14. ~~A~~ Example of a questionnaire that ~~about program activities.~~ ~~Also~~ uses both closed and open response formats.



DOCUMENTATION QUESTIONNAIRE
Peer-Tutoring Program

Please answer the following questions by placing the letter of the most accurate response on the line to the left of the question. We are interested in finding out what the project looked like in operation during the past week, regardless of how it was planned to look. If more than one answer is true, answer with as many letters as you need.

____ 1. On the average, how many times did tutoring sessions take place in your classroom?

a) never
b) 1 or 2 times
c) 3 or 4 times
d) 5 or more times

____ 2. What was the average length of a tutoring session?

a) 5-15 minutes
b) 15-25 minutes
c) 25-45 minutes
d) longer than 45 minutes

____ 3. Where in the school did tutoring usually take place?

a) classroom
b) sometimes classroom, sometimes other room
c) library
d) room other than classroom or library

____ 4. Who were the tutors?

a) only fast students
b) fast students and some average students
c) only average students
d) other

____ 5. On what basis were tutees selected?

a) reading achievement
b) teacher recommendations
c) general grade average
d) other

____ 6. What materials were used by teachers and tutors?

a) whatever tutors chose
b) specially constructed games
c) "Read and Say" workbooks
d) other

____ 7. How typical of the program as a whole was last week, as you have described it here?

a) just the same
b) almost the same
c) some aspects not typical
d) not typical at all

Figure 15. Example of a closed response questionnaire

Outline for Chapter 6

METHODS FOR MEASURING PROGRAM IMPLEMENTATION: OBSERVATIONS

A. Introduction- Setting Up an Observation System

1. Range of possibilities in observation/participant observation "systems"
    - a. Informal, casual, seat of the pants
    - b. More credible "scientific"/systematic approaches ranging along continuum from highly prestructured to those based on emerging information
        - 1) Quantitative, highly structured, predetermined categories, "research"
        - 2) Qualitative, participant observation, categories emerge from analysis of field notes, move back and forth from data to analysis

2. Note limitation when people think they're doing qualitative study when in fact they're using a casual, unsystematic approach; cite Patton's kit book

B. Making Quantitative Observations-- Steps 1-12 (pp. 90-112); plus Stallings Observation System (pp. 112-115)-- editorial changes only

C. Making Qualitative Observations (Each step will be elaborated; examples added as necessary for clarification)

1. Step 1. Construct a program characteristics list describing what the program should look like

2. Step 2. Make initial contact with program personnel, conduct initial observations, establish entree and rapport, inform program staff about participant/observation

3. Step 3. Develop an evaluation timeline based on analysis of your initial information, prepare a "sampling plan" for observations, decide how much time can be spent doing observations, if possible, write out the program "theory"

4. Step 4. Assemble the evaluation team (people familiar with naturalistic methods), decide on appropriate format for fieldnotes, discuss evaluation context, initial "findings," critical issues; arrange analysis schedule

5. Step 5. Move back and forth between collecting data (from participant/observation, interviews, questionnaires, i.e., whatever data collection techniques are appropriate) AND analyzing data
    - a. Do this until you have sufficient information to answer the users' questions (or you run out of time)

- b. Part of process is a series of meetings to discuss themes, issues, critical incidents; these can involve evaluators and program personnel as appropriate
- c. "Thick description" should be written during the process; ongoing evolution of written description of program; should be given to program people for reaction, then revised as need be

6. Step 6. When all data are in, prepare them for interpretation and final presentation

D. Chapter Summary
   1. Importance of observation techniques
   2. Selection of appropriate level of "rigor" in observations as well as appropriate type

E. (Updated) For Further Reading (many texts now available)

Appendix

AN OUTLINE OF AN IMPLEMENTATION REPORT

The outline in this appendix will yield a report describing program implementation only. In most evaluations, implementation issues comprise only one facet of a more elaborate enterprise concerned with the design of the evaluation, the intended outcomes of the program, the measures used to assess achievement of those outcomes, and the results these measures produced. If this description of an extended evaluation responsibility matches your task, then you will need to incorporate information from the outline here into a larger report discussing other aspects of the program and its evaluation. If, in fact, the evaluation compares the effects of two different programs considered equally important by your audience, then you should prepare an implementation report to describe them both.

The headings in this appendix are organized according to the ~~five~~ six major sections of an implementation report:

1. A summary that gives the reader a quick synopsis of the report
2. A description of the context in which the program has been implemented, focusing mainly on the setting, administrative arrangements, personnel, and resources involved
3. A description of the point of view from which implementation has been examined. This section can have one of two characters:

   a. It can describe the program's most critical features as prescribed by a program plan, a theory or teaching model, or someone's predictions about what will make the program succeed or fail, or

   b. It can explain the qualitative evaluator's choice not to use a prescription to guide her examination of the program.

4. A description of the implementation evaluation itself--the choice of measures, the range of program activities examined, the sites examined, and so forth. This section also includes a rationale for choosing the data sources listed.

5. Results of implementation backup measures and discussions of program implementation. This section can do one of two things:

   a. Describe the extent to which the program as implemented fit the one that was planned or prescribed by a plan, theory, or teaching model

   b. Describe implementation independent of underlying intent. This description, usually gathered using a naturalistic method, reflects a decisions that the evaluator describe what she discovered rather than compare program events with underlying points of view.

   In either case, this section describes what has been found, noting variations in the program across sites or time.

6. Interpretation of results, commendations, and suggestions for further program development or evaluation.

Report Section 1. Summary

The summary is a brief overview of the report, explaining why a description of implementation has been undertaken and listing the major conclusions and recommendations to be found in Section 6. Since the summary is designed for people who are too busy to read the full report, it should be limited to one or two pages, maximum. Although the summary is placed first in the report, it is the last section to be written.

Report Section 2. Background and Context of the Program

This section sets the program in context. It describes how the program was initiated, what it was supposed to do, and the resources available. The amount of information presented will depend upon the audiences for whom the report has been prepared. If the audience has no knowledge of the program, the program must be fully described. If, on the other hand, the implementation report is mainly intended for internal use and its readers are likely to be familiar with the program, this section can be brief and set down information "for the record." Regardless of the audience, if your report will be written, it might become the only lasting record of the program's implementation. In this case, the context section should contain considerable data.

¶ If your program's setting includes many different schools or districts, it may not be practical to cover every evaluation issue separately for each school or program site. Instead, for each issue indicate similarities and differences among schools or sites or the range represented or the most typical pattern that occurred.

Report Section 3. General Description of the Critical Features of the Program as Planned--Materials and Activities